FiboSeg: Fully automated segmentation of upper and lower jaws from 3D intra-oral surface scanners

Mathieu Leclercq MATHIEU.LECLERCQ@CPE.FR
Antonio Ruellas ARUELLAS@UMICH.EDU
Marcela Gurgel MLIMAGUR@UMICH.EDU
Marilia Yatabe MSYATABE@UMICH.EDU
Jonas Bianchi BIANCHIJ@UMICH.EDU
Lucia Cevidanes LUCIACEV@UMICH.EDU
Martin Styner STYNER@CS.UNC.EDU
Beatriz Paniagua BEATRIZ.PANIAGUA@KITWARE.COM
Juan Carlos Prieto JPRIETO@MED.UNC.EDU

Editors: Under Review for MIDL 2022

Abstract

In this paper, we present a deep learning based method for surface segmentation. This technique consists of acquiring 2D views and extracting features from the surface such as the normal vectors. The rendered images are analyzed with a 2D convolutional neural network, such as UNETs. We test our method in a dental application for segmentation of dental crowns. The neural network is trained for the multi-class segmentation, using image labels as ground truth. The segmentation task achieved an average Dice of 0.97, sensitivity of 0.97 and precision of 0.97.

Keywords: Deep Learning, segmentation, dental crown, Universal label id, intra-oral surface

1. Introduction

Developments in dentistry have resulted in an improved application of 3D technologies such as intra-oral surface (IOS) scanners which are used to design ceramic crowns, veneers, inlays, and occlusal guards, as well as assisting with implants. IOS are being used more often for automated diagnosis such as caries detection (Lim et al., 2021), analysis of risk factors of tooth movement(Commer et al., 2000), and treatment planning(Flügge et al., 2013). These 3D surface models require shape analysis techniques for analyzing and understanding the geometry and achieve state of the art performance for tasks such as segmentation, classification, and retrieval. In this paper we present a novel method for 3D surface segmentation based on a multi-view approach. Fast and accurate segmentation of the IOS remains a challenge due to various geometrical shapes of teeth, complex tooth arrangements, different dental model qualities, and varying degrees of crowding problems(Li and Wang, 2016). Our target application is the multi-class segmentation following the Universal Numbering System proposed by the American Dental Asociation (ADA), the dental notation system used in the United States.

The multi-view approach consists of generating 2D images of the 3D surface from different view points. The generated images serve as a training set for a neural network. We use Pytorch3D[1] to generate images on the fly during training and a one-to-one mapping that relates faces in the 3D model and pixels in the generated images. This is useful in inference time when we have to put the resulting labels from the images back into the 3D model. The remainder of the manuscript is organized as follow: the materials used in this study, related work, details of our implementation, results and conclusions.

2. Materials

The dataset consists of 78 mandibular IOS (40 for the upper and 38 for the lower dental arches). Digital dental model of the mandibular arch was acquired from intra oral scanning with the TRIOS 3D intra oral scanner. The TRIOS intra oral scanner (IOS) utilizes "ultrafast optical sectioning" and confocal microscopy to generate 3D images from multiple 2D images with an accuracy of 6.9 $\pm$ 0.9 μ m. All scans were obtained according to the manufacturer's instructions, by one trained operator. The training was done on an NVIDIA TITAN V GPU with 12GB of memory.

3. Related work

3.1. 3D shape analysis

Learning-based methods for shape analysis use the 3D models to learn descriptors directly from them, however, adapting state-of-the art 2D CNNs to work on 3D shapes is a challenging task. The main impediment is the arbitrary structures of 3D models which are usually represented by point clouds or triangular meshes, whereas deep learning algorithms use the regular grid-like structures found in 2D/3D images(Boubolo et al., 2021; Deleat-Besson et al., 2021).

Multi-view approaches adapt 3D objects by rendering them from different view-points and use the snapshots to extract 2D image features using 2D CNNs(Su et al., 2015; Kanezaki et al., 2018; Ma et al., 2018). On the other hand, volumetric approaches use 3D voxel grids to represent the shape and apply 3D convolutions to learn shape features(Wu et al., 2015; Wang et al., 2017; Riegler et al., 2017). Finally, other approaches consume the point clouds directly and implement multi-layer-perceptrons and/or transformer architectures, or a generalization of typical CNNs (Qi et al., 2017a; Lian et al., 2020; Li et al., 2018; Wu et al., 2019).

3.2. Intra oral surface segmentation

MeshSegNet(Lian et al., 2020) uses raw surface attributes as inputs and integrates graph-constrained learning modules then, a dense fusion strategy is applied to combine local-to-global geometric features for the learning of higher-level features for mesh cell annotation. The predictions by MeshSegNet are further post-processed by a graph-cut refinement step for final segmentation. "Deep Learning Approach to Semantic Segmentation in 3D Point Cloud Intra-oral Scans of Teeth" (Zanjani et al., 2019) proposes an end-to-end deep learning

1. https://pytorch3d.org/

framework for segmentation of teeth from point clouds representing IOS. TSegNet (Cui et al., 2021) developed a fully automatic algorithm to segment tooth on 3D dental models guided by the tooth centroid information. Mask-MCNet(Zanjani et al., 2021) localizes each individual tooth instance by predicting its 3D bounding box and segments the points that belong to each individual tooth instance. FlyByCNN(Boubolo et al., 2021; Deleat-Besson et al., 2021) is a multi-view based approach that uses UNETs(Ronneberger et al., 2015) to segment each individual image. The merging and annotating teeth and root canals approach (Deleat-Besson et al., 2021) uses a classification model to identify upper or lower jaws. Then, it proceeds to align the 3D objects to a template and then label each crown with the universal id.

4. Method

4.1. Rendering the 2D views

The Pytorch3D framework allows rendering the 3D object from different view points and extract views that can be fed to a CNN in a end-to-end training procedure. The rendering engine provides a map that relates pixels in the images to faces in the mesh and allows rapid extraction of point data (normals, curvatures, labels, etc.) as well as setting information back into the mesh after inference. In order to get different viewpoints, we apply random rotations to the camera, so that it moves on the surface of a unit sphere. For each snapshot, we generate two images. The first one contains the surfaces normals encoded in the RGB components. The second one are the label maps that are used as ground truth in the segmentation task. We set the resolution of the rendered images to 320px. We use ambient lights so that the rendered images don't have any specular components.

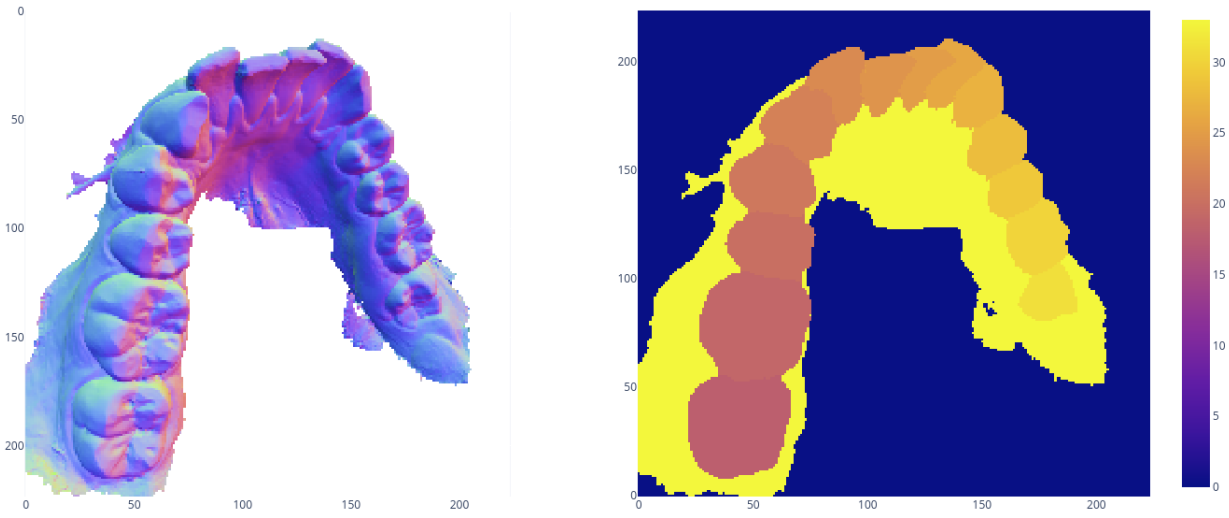


Figure 1: Example of a rendered 2D view. Left: surface normals encoded in RGB components, (additional surface properties may be rendered or extracted from the surface using the face-id maps). Right: ground truth labels for the dental crowns rendered with a color map. These image pairs are used in training of a segmentation task.

4.2. Training the network

We have in total 78 different IOS. We use 55 for the training, 7 for validation, and 16 for testing. To create the neural networks we use MONAI[2], an open-source Pytorch-based framework for deep learning in healthcare imaging. MONAI provides a complete framework to easily create Datasets and integrate neural networks such as a UNET. We use the DiceCELoss, which computes the Dice Loss as well as Cross-Entropy Loss and returns the weighted sum of these two. This loss is frequently used for segmentation tasks. The learning rate is set to $1e^{-}4$ using the Adam optimizer. One important thing to note is that there is no previous pre-processing to the mesh, sub-sampling of points/faces or any classification task to identify upper or lower jaws. The training learns to identify 34 different labels corresponding to the upper and lower crowns. We use one-hot encoding for the 34 different classes: 32 different crowns, in addition to the gum and the background. To make the validation more accurate, we implemented a 5-fold cross-validation. The results for this cross-validation can be found in the appendix.

4.3. Prediction

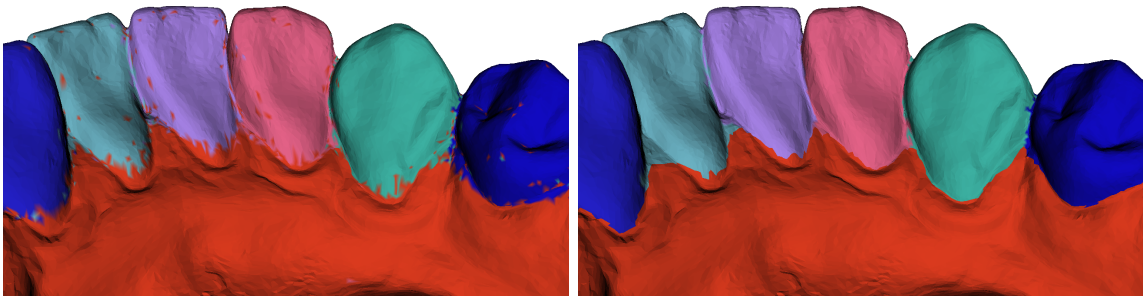
The prediction is composed of three major steps: 1. Render 2D views from the 3D object; 2. Run inference on the 2D views. 3. map the information back into the 3D mesh. We render

2. monai.io

70 views with a resolution of 320px. The position of the cameras are distributed on the surface of sphere and follow a Fibonacci lattice. This method ensures an even distribution of the viewpoints on the surface of a sphere, and thus an efficient prediction as we capture the object regularly. After we run inference on the 2D views, we use a weighted majority voting scheme to put information back into the 3D mesh. The Pytorch3d rasterizer returns a mapping that keeps track of the nearest face at each pixel, for each 2D image that is created (pixels that are hit by zero face are padded with -1). We can use this variable to give at each iteration a label to every face that has been assigned to a pixel. As there are 34 different classes, the outputs of the network are of shape $[1, \text{img\_size}, \text{img\_size}, 34]$. After a few iterations, almost every face has been assigned to one or more outputs. When all the snapshots have been generated we can proceed to the weighted majority voting; the weights being the output of the network to which we apply the softmax function.

4.4. Post-Process

In the event that some faces of the surface are not assigned to any output at the end of the prediction, we apply an 'island removal' approach, that assigns the closest-connected label. Figure 4.4 shows an example of the island removal approach. Unlabeled points in the crowns get assigned the closest label.



(a) (b)

Figure 2: Island remove approach. Unlabeled points are assigned the closest-connected label.

4.5. Distance metric

One way to tell if our model is accurate is to compute the average distance between the borders of the ground truth and the predicted labels for every crown. We do this by isolating each crown with a threshold and then extracting the points on the border. We then compute the distance between each border point on the predicted label and the closest border point on the ground truth. Finally, we divide the sum of the distances for every point by the number of points to get the average distance for one specific crown.

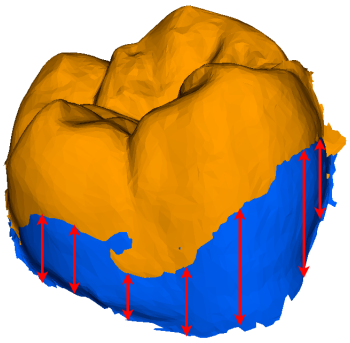
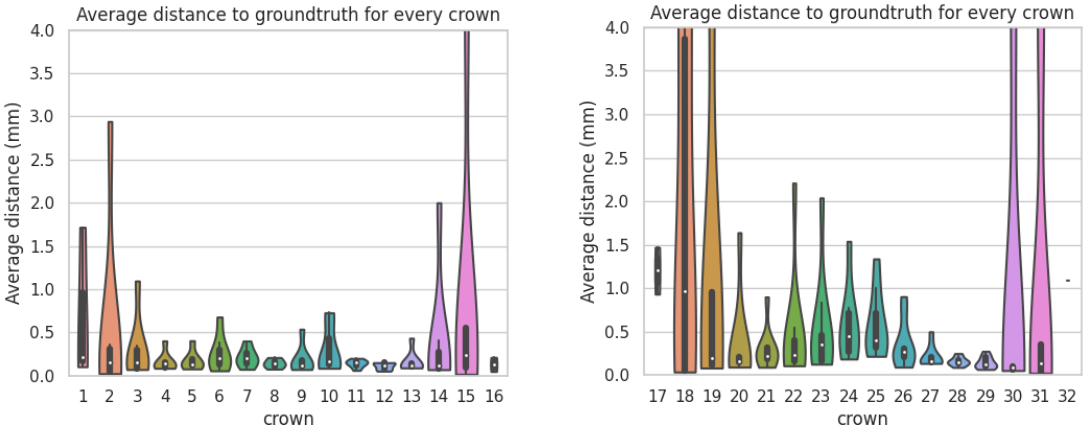


Figure 3: Predicted label (orange) on top of ground truth (blue)

5. Results

5.1. Distance metric



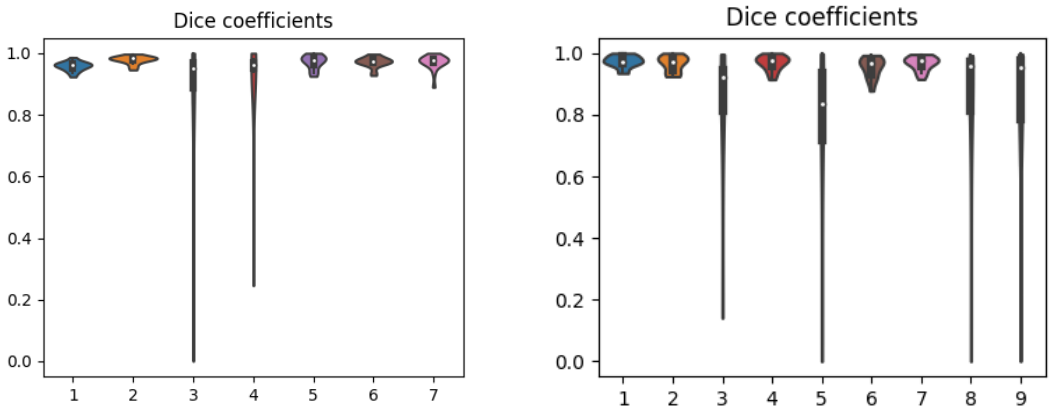
(a) (b)

Figure 4: Violin plots showing the distance metric. (a) is upper jaw and (b) is lower jaw.

The better results for the upper jaw can be explained by the fact that the test set for the lower jaw contains more models with wisdom teeth (4 out of the 9 models in total). The model was trained with a majority of surfaces with only 14 crowns and no wisdom teeth. The poor results for the labels that are just next to the wisdom teeth come from the fact that most wisdom teeth got the label from the next crown. An incorrect labeling of some crowns tends to decrease the accuracy of the prediction for the other crowns. Further results show this phenomenon quite well.

5.2. Dice coefficients

5 out of the 7 IOS for the upper jaw have Dice coefficients > 0.9. On these 5 scans all crowns are correctly labeled and the result is almost identical to the ground truth. The



(a) (b)

Figure 5: Dice coefficients. (a) is upper jaw and (b) is lower. In each plot, each "violin" represents a separate 3D surface from the test set. Samples with wisdom teeth get worse Dice coefficients because the model was trained with a great majority of models with only 14 teeth.

worse result for the 3^{rd} scan may come from the fact that the IOS looks different from the training data (see Appendix). The 4^{th} scan for the upper jaw and the 3^{rd}, 5^{th}, 8^{th} and 9^{th} for the lower jaw show worse Dice coefficients because they contain 16 teeth (wisdom teeth). As mentioned before, the training set contains a very few number of samples with wisdom teeth. With a wider range of training data we can expect much better results for these types of scans as well.

Figure 6 shows that when the network makes predictions on data types for which it has trained on, it produces excellent results. The Dice coefficients are always above 0.9 and even above 0.95 for 25 out of the 28 different crowns.

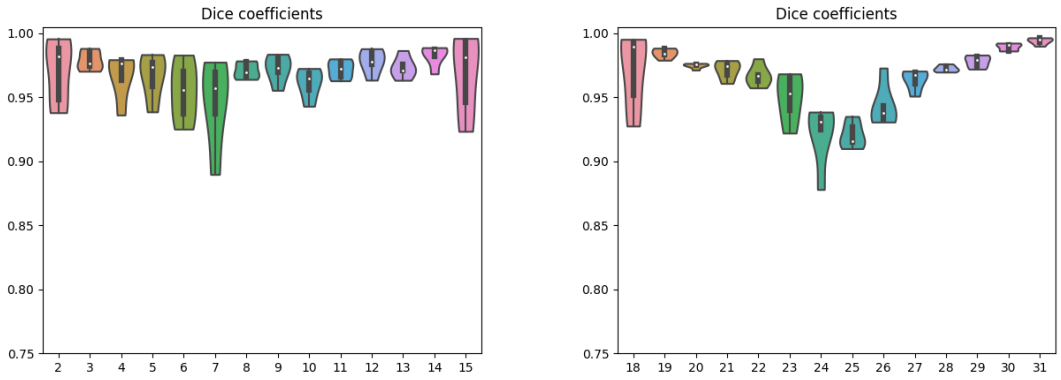
5.3. Resulting labeled IOS

Figure 7 shows the output segmentation for one of our test cases.

5.4. Comparison with competing methods

Metric	PointNet	PointNet++	PointConv	MeshSegNet	Our method
DSC	0.840 ± 0.117	0.907 ± 0.067	0.939 ± 0.060	$\mathbf{0.981 \pm 0.028}$	0.968
SEN	0.910 ± 0.129	0.988 ± 0.037	0.983 ± 0.047	$\mathbf{0.989 \pm 0.035}$	0.97
PPV	0.795 ± 0.132	0.846 ± 0.107	0.904 ± 0.086	$\mathbf{0.975 \pm 0.034}$	**0.97**

This table shows results for 3 state-of-the-art deep learning methods ((Qi et al., 2017a), (Qi et al., 2017b) and (Wu et al., 2019)) as well as (Lian et al., 2020). The results for our



(a) (b)

Figure 6: Dice coefficients for jaws with no wisdom teeth. (a) is upper jaw and (b) is lower. We picked from the test set 5 upper jaws and 5 lower jaws with just the 14 "regular" teeth (i.e. no wisdom teeth). For each plot, each "violin" represents a separate crown.

segmentation method were obtained using the data set samples that do not have wisdom teeth.

6. Conclusion

This new method for automatic multi-class segmentation of 3D surfaces has proven to be accurate and effective, as well as easy to integrate into existing pipelines. The surfaces models can be directly processed by the approach and no intermediate pre-processing or down-sampling steps for the surface model are needed. A great advantage of this method is the ability to predict the universal ids of the crowns in the upper and lower jaws. The competing approaches focused on upper models only or use a classification model to identify upper/lower jaws. Our approach is fully automated and labels both upper and lower crowns with a single model. The model learns to identify features specific to each jaw. We are aware that this model can still be improved with larger and more diverse data sets. However, the results obtained are competitive with existing methods such as MeshSegNet and PointConv. Prediction for jaws with no wisdom teeth is excellent, and we can safely expect better results for wisdom teeth segmentation as our sample size increases.

References

Louis Boubolo, Maxime Dumont, Serge Brosset, Jonas Bianchi, Antonio Ruellas, Marcela Gurgel, Camila Massaro, Aron Aliaga Del Castillo, Marcos Ioshida, Marilia S Yatabe, et al. Flyby cnn: a 3d surface segmentation framework. In *Medical Imaging 2021: Image Processing*, volume 11596, page 115962B. International Society for Optics and Photonics, 2021.

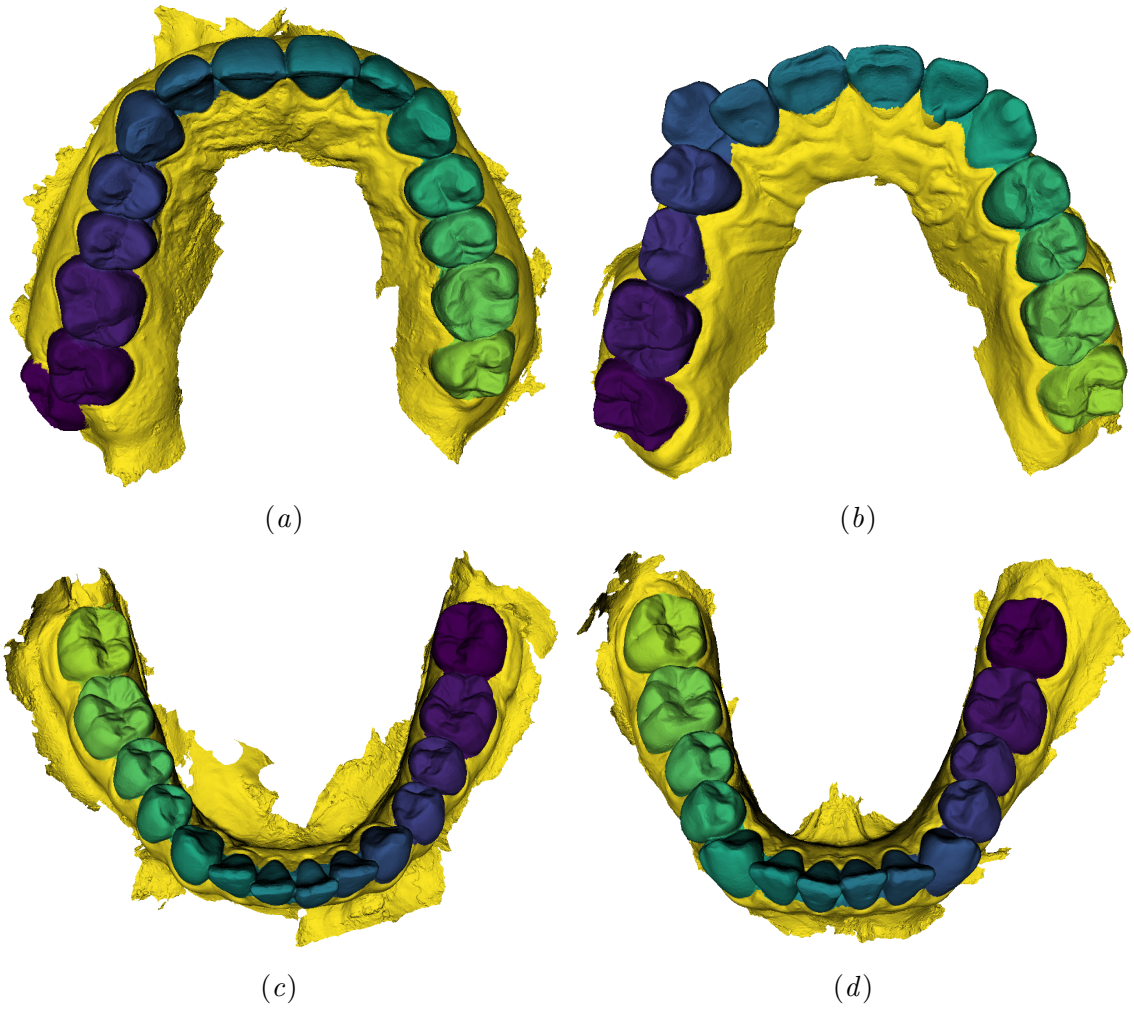


Figure 7: Resulting labeled IOS from the test set. (a) and (b) are upper jaws. (c) and (d) are lower jaws.

P Commer, C Bourauel, K Maier, and A Jäger. Construction and testing of a computer-based intraoral laser scanner for determining tooth positions. *Medical engineering & physics*, 22(9):625–635, 2000.

Zhiming Cui, Changjian Li, Nenglun Chen, Guodong Wei, Runnan Chen, Yuanfeng Zhou, Dinggang Shen, and Wenping Wang. Tsegnet: an efficient and accurate tooth segmentation network on 3d dental model. *Medical Image Analysis*, 69:101949, 2021.

Romain Deleat-Besson, Celia Le, Winston Zhang, Najla Al Turkestani, Lucia Cevidanes, Jonas Bianchi, Antonio Ruellas, Marcela Gurgel, Camila Massaro, Aron Aliaga Del Castillo, et al. Merging and annotating teeth and roots from automated segmentation

of multimodal images. In *International Workshop on Multimodal Learning for Clinical Decision Support*, pages 81–92. Springer, 2021.

Tabea V Flügge, Stefan Schlager, Katja Nelson, Susanne Nahles, and Marc C Metzger. Precision of intraoral digital dental impressions with itero and extraoral digitization with the itero and a model scanner. *American journal of orthodontics and dentofacial orthopedics*, 144(3):471–478, 2013.

Asako Kanezaki, Yasuyuki Matsushita, and Yoshifumi Nishida. Rotationnet: Joint object categorization and pose estimation using multiviews from unsupervised viewpoints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5010–5019, 2018.

Yangyan Li, Rui Bu, Mingchao Sun, Wei Wu, Xinhan Di, and Baoquan Chen. Pointcnn: Convolution on x-transformed points. *Advances in neural information processing systems*, 31, 2018.

Zhongyi Li and Hao Wang. Interactive tooth separation from dental model using segmentation field. *PloS one*, 11(8):e0161159, 2016.

Chunfeng Lian, Li Wang, Tai-Hsien Wu, Fan Wang, Pew-Thian Yap, Ching-Chang Ko, and Dinggang Shen. Deep multi-scale mesh feature learning for automated labeling of raw dental surfaces from 3d intraoral scanners. *IEEE transactions on medical imaging*, 39(7): 2440–2450, 2020.

Jung-Hwa Lim, Utkarsh Mangal, Na-Eun Nam, Sung-Hwan Choi, June-Sung Shim, and Jong-Eun Kim. A comparison of accuracy of different dental restorative materials between intraoral scanning and conventional impression-taking: An in vitro study. *Materials*, 14 (8):2060, 2021.

Chao Ma, Yulan Guo, Jungang Yang, and Wei An. Learning multi-view representation with lstm for 3-d shape recognition and retrieval. *IEEE Transactions on Multimedia*, 21(5): 1169–1182, 2018.

Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017a.

Charles R Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *arXiv preprint arXiv:1706.02413*, 2017b.

Gernot Riegler, Ali Osman Ulusoy, and Andreas Geiger. Octnet: Learning deep 3d representations at high resolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3577–3586, 2017.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 945–953, 2015.

Peng-Shuai Wang, Yang Liu, Yu-Xiao Guo, Chun-Yu Sun, and Xin Tong. O-cnn: Octree-based convolutional neural networks for 3d shape analysis. *ACM Transactions on Graphics (TOG)*, 36(4):1–11, 2017.

Wenxuan Wu, Zhongang Qi, and Li Fuxin. Pointconv: Deep convolutional networks on 3d point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9621–9630, 2019.

Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1912–1920, 2015.

Farhad Ghazvinian Zanjani, David Anssari Moin, Bas Verheij, Frank Claessen, Teo Cherici, Tao Tan, et al. Deep learning approach to semantic segmentation in 3d point cloud intra-oral scans of teeth. In *International Conference on Medical Imaging with Deep Learning*, pages 557–571. PMLR, 2019.

Farhad Ghazvinian Zanjani, Arash Pourtaherian, Svitlana Zinger, David Anssari Moin, Frank Claessen, Teo Cherici, Sarah Parinussa, and Peter HN de With. Mask-mcnet: Tooth instance segmentation in 3d point clouds of intra-oral scans. *Neurocomputing*, 453: 286–298, 2021.

Appendix A. Confusion matrices

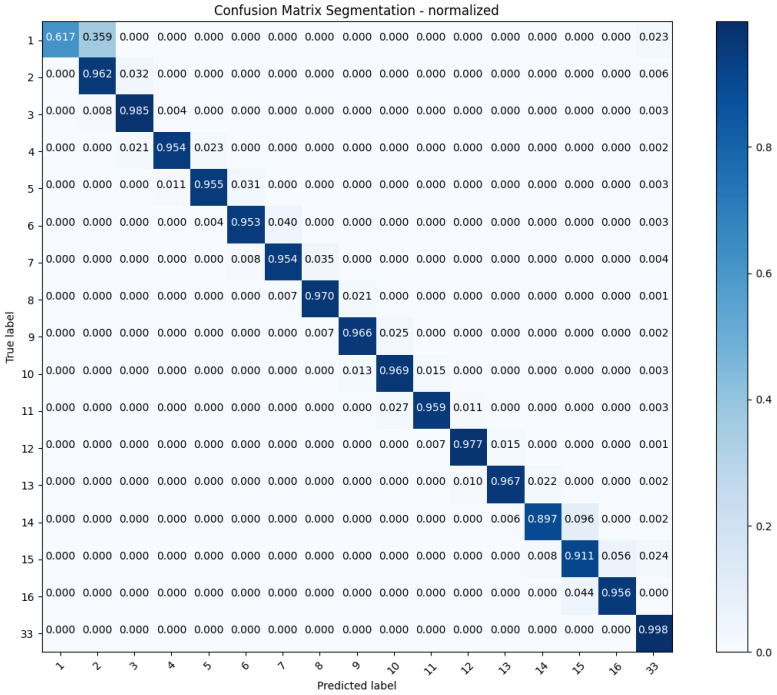


Figure 8: Confusion matrix for the upper jaw

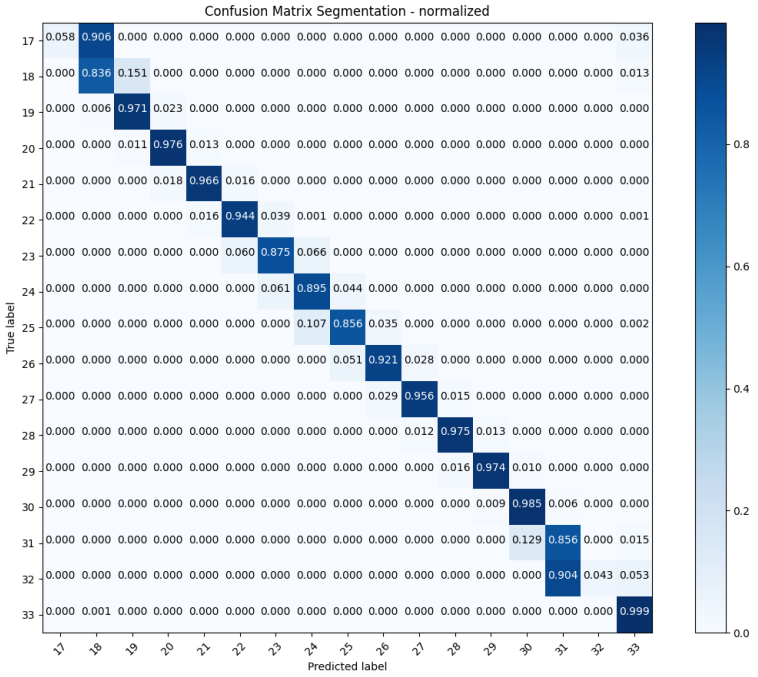
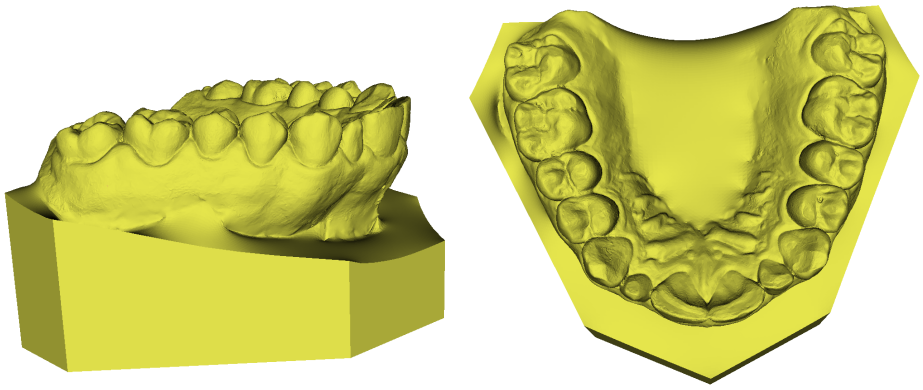


Figure 9: Confusion matrix for the lower jaw

These confusion matrices show that the neural network gives better results for upper wisdom teeth than for the lower ones. This difference comes from the nature of the training set: more lower jaws models with wisdom teeth were put into the test set, so there were fewer ones remaining for the training set. This result is rather positive and shows that the model can be improved with a better training set.

Appendix B. Detailed results for test samples with no wisdom teeth

Label	PPV	SEN	f1-score	support
2	0.97	0.98	0.97	17885
3	0.97	0.98	0.98	23869
4	0.98	0.96	0.97	14613
5	0.98	0.95	0.97	16047
6	0.95	0.95	0.95	12707
7	0.94	0.96	0.95	10626
8	0.97	0.98	0.97	15119
9	0.97	0.98	0.97	15596
10	0.95	0.97	0.96	11149
11	0.98	0.96	0.97	13454
12	0.98	0.98	0.98	15783
13	0.97	0.97	0.97	14924
14	0.98	0.98	0.98	23135
15	0.98	0.96	0.97	17584
16	0.99	0.99	0.99	2126
18	0.97	0.98	0.97	21550
19	0.99	0.98	0.98	24258
20	0.97	0.98	0.97	14493
21	0.98	0.97	0.97	13409
22	0.98	0.96	0.97	12153
23	0.95	0.95	0.95	9343
24	0.91	0.94	0.92	9024
25	0.93	0.91	0.92	8708
26	0.96	0.93	0.94	9593
27	0.96	0.97	0.96	12704
28	0.97	0.98	0.97	13177
29	0.98	0.98	0.98	15131
30	0.99	0.99	0.99	25318
31	0.99	1.00	0.99	21575
33	1.00	1.00	1.00	540485
avg	0.97	0.97	0.97	

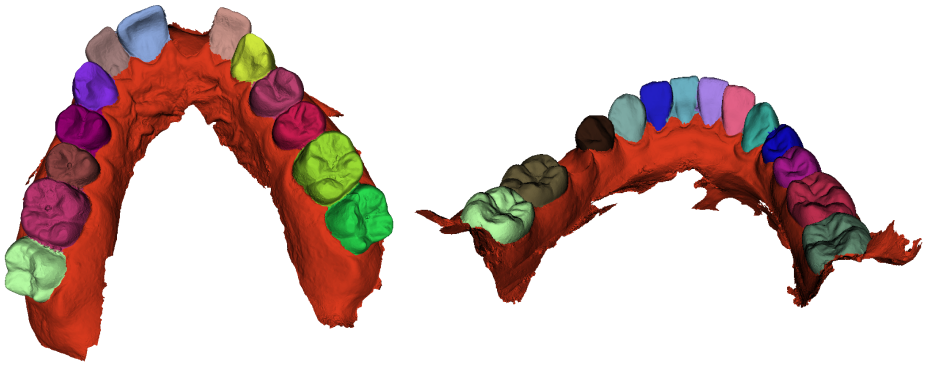


(a) (b)

Figure 10: 3^{rd} upper jaw in the test set. It is the only IOS with this kind of base. This may explain why the neural network struggles to produce very accurate results for this particular model.

Appendix C. Third upper jaw in the test set

Appendix D. Prediction with random teeth removal



(a) (b)

Figure 11: This figure shows two examples of predictions in which a crown was randomly removed from the surface. The prediction is accurate. This example does not exactly reproduce the case in which a crown is naturally missing from the jaw, because of the gap that it creates. We did not have such cases in our data set.

Appendix E. 5-fold cross-validation results on samples with 14 teeth (no wisdom teeth)

Metric	Cross-Validation
Precision	0.98 ± 0.03
Sensitivity	0.98 ± 0.11
f1-score	0.98 ± 0.08
Dice	0.96 ± 0.03

This table shows the results for the 5-fold cross-validation. Dice is at 0.96, which is slightly lower than the 0.97 value obtained with the first training. We had to stop the trainings manually without relying on the early stopping criteria in order to have them all done on time. This could explain the slight difference. These results show that our method provides accurate results in the case of jaws with 14 teeth, even on a wider test set.