Clinically Applicable System For 3D Teeth Segmentation in Intraoral Scans using Deep Learning

Jin Hao
Harvard University

Wen Liao
Sichuan University

Yueling Zhang
Sichuan University

Peilin Li
Sichuan University

Jianru Yi
Sichuan University

Jerry Peng
DeepAlign Tech Inc.

Zeu Zhao
DeepAlign Tech Inc.

Zhang Chen
DeepAlign Tech Inc.

Wenxuan Shi
DeepAlign Tech Inc.

Tingyu Chen
DeepAlign Tech Inc.

Bowen Zhou
Angel Align Inc.

Yang Feng
Angel Align Inc.

Bing Fang
Department of Orthodontics, Shanghai Ninth People's Hospital, Collage of Stomatology, Shanghai Jiao Tong University School of Medicine

Haoji Hu
Zhejiang University

Howard Yang
Zhejiang University

Erping Li
Zhejiang University https://orcid.org/0000-0002-5006-7399

Zuozhu Liu
Zhejiang University

Zhihe Zhao (✉ zhzhao@scu.edu.cn)
Sichuan University

Article

Keywords: 3D Teeth Segmentation, Intraoral Scans, Deep Learning, Digital Dentistry

Posted Date: November 10th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-103285/v1

Clinically Applicable System For 3D Teeth Segmentation in Intraoral Scans using Deep Learning

Jin Hao[1, 2], Wen Liao[1], Yueling Zhang[1], Peilin Li[1], Jianru Yi[1], Jerry Peng[3], Zeu Zhao[3], Zhang Chen[3], Wenxuan Shi[3], Tingyu Chen[3], Bowen Zhou[4], Yang Feng[4], Bing Fang[5], Haoji Hu[6], Howard H. Yang[7], Erping Li[6, 7], Zuozhu Liu[6, 7*], Zhihe Zhao[1*]

[1] State Key Laboratory of Oral Diseases & National Clinical Research Center for Oral Diseases & West China Hospital of Stomatology, Sichuan University, Chengdu, China
[2] Harvard School of Dental Medicine, Harvard University, Boston, MA, USA
[3] DeepAlign Tech Inc., Ningbo, China
[4] Angel Align Inc., Shanghai, China
[5] Ninth People's Hospital Affiliated to Shanghai Jiao Tong University, Shanghai Research Institute of Stomatology, National Clinical Research Center of Stomatology, Shanghai, China
[6] College of Information Science and Electrical Engineering, Zhejiang University, Hangzhou, China
[7] Zhejiang University-University of Illinois at Urbana-Champaign Institute, Zhejiang University, Haining, China

*Correspondence: zuozhuliu@intl.zju.edu.cn (Z.L.) & zhzhao@scu.edu.cn (Z.Z.)

Abstract

Digital dentistry plays a pivotal role in dental healthcare. A critical step in many digital dental systems is to accurately delineate individual teeth and the gingiva in the three-dimension (3D) intraoral scanned (IOS) mesh data. However, previous state-of-the-art methods are either time-consuming or error-prone, hence hinder their clinical applicability. In this paper, we present an accurate, efficient, and fully-automated deep learning model, trained on a dataset of 4,000 IOS data annotated by experienced human experts. On a hold-out dataset of 200 scans, our model achieves a per-face accuracy, average-area accuracy and area under the receiver operating characteristic curve (AUC) of 96.94%, 98.26%, and 0.9991, respectively, significantly outperforming the state-of-the-art baseline. In addition, our model only takes about 24 seconds to generate segmentation outputs, as compared to over 5 minutes by the baseline and 15 minutes by human experts. A clinical performance test of 500 patients with malocclusion or/and abnormal teeth shows that 96.9% of the segmentations are satisfactory for clinical applications, 2.9% automatically trigger alarms for human improvement, and only 0.2% of them need rework. Our research demonstrates the potential for deep learning to improve the efficacy and efficiency of dental treatment and digital dentistry.

INTRODUCTION

With the increasing demands for dental healthcare, computer-aided (CAD) dental systems or digital dentistry is becoming a fundamental component in both dental industry and research. Alongside the treatment of many oral diseases, such as orthodontics and implant, the first crucial step is to precisely recognize individual teeth and the gingiva in the 3D intra-oral scanned (IOS) teeth data collected from patients [1]. Formally, given an IOS 3D teeth scan, which is usually denoted as a mesh consisting of triangulated faces, teeth segmentation aims to classify each face into different teeth and the gingiva following the Federation Dentaire Internationale (FDI) standard [2]. A segmentation with high accuracy and resolution is an indispensable prerequisite for subsequent clinical applications. For example, dentists will remove or re-align (i.e., move and rotate) segmented teeth in the CAD software to simulate orthodontic or implant treatment. Moreover, the segmentation system also requires high automation and efficiency owing to conflicts between the large number of patients and the limited labor and computational resources in practice.

The semantic segmentation over the teeth and gingiva of the 3D IOS mesh data remains a very challenging task, while an accurate, efficient, and fully-automated clinically applicable solution is yet under development. The key challenges are as follows. First, unlike other common medical data (e.g., breast or lung cancer MRI images), the IOS teeth data varies significantly among patients. For example, the number of teeth (e.g., hypodontia or hyperdontia) and the dental arch forms (e.g., tapering or squared or ovoid arch) are not consistent across different patients, not to mention other more complicated anatomy such as positional varieties (i.e., tipped, rotated or shifted tooth, crowding teeth), tooth shape alteration (e.g., tooth with cavities or defective restoration or attrition), and tooth size abnormalities (e.g., microdontia, macrodontia). A fully-automated system is expected to take all of these issues into account and provide robust and reliable segmentation outputs. Second, a clinically applicable system needs to generate fine-grained segmentations for high-resolution data, while the teeth-teeth and teeth-gingival boundaries are usually indistinguishable and tightly fused. For example, the teeth-gingiva boundaries from patients with gingival recession are slightly different from normal ones. Failing to recognize such minor differences can lead to severe repercussions in the subsequent treatment planning. Finally, a short processing time of the segmentation system is also of necessity to provide real-time feedback to support clinical diagnosis, which could substantially improve the throughput and quality of services in clinics.

Recently, a plethora of existing works have launched attempts to address these challenges. Traditional geometry-based methods extract hand-crafted features such as curvatures from the IOS mesh data to design decision rules for segmentation [3-10]. Though with agreeable interpretability, these methods cannot generalize to various teeth morphologies and usually require human involvement for post-correction. Recent advances employed deep learning techniques to develop more generic solutions [8-14]. Particularly, by leveraging the convolutional neural networks (CNNs), one group of researchers segmented the mesh data with predefined geometrical features of mesh faces [11], while another group employs a hierarchical strategy to synthesize gingiva, incisors, molars and premolars step-by-step [12]. These methods showed better performance than geometry-based approaches. However, they are limited in scope and/or scale, with unsatisfactory results in complicated cases and performance tested only with a limited number of samples. More importantly, previous works have not evaluated the clinical utility and applicability of their models in patients with severe malocclusion and/or dental defects, and to what extent these models could help the industry and clinics.

To cope with these limitations, we propose an accurate, efficient, fully-automated and fault-aware system based on deep learning for clinically applicable teeth segmentation. The entire system, which is referred to as DC-Net (Figure 1), contains two modules: the **D**eep learning module and the **C**anary module. The deep learning module is to generate fine-grained segmentation outputs, while the canary module is to evaluate whether these outputs are accurate for clinical use or manual corrections are required. In lieu of directly operating on mesh scans, we transform the mesh data into point clouds, which maintain the original geometric structures but are much easier to process with neural networks [15-17]. To learn meaningful representations from high-resolution point clouds, we employ the Edge-conv block as the cornerstone in the deep neural network [15]. The Edge-conv block is designed to aggregate local semantic geometric information, and global characteristics can be synthesized by stacking multiple Edge-conv blocks. The segmented results from the deep neural network are further refined with a graph-cut algorithm to improve the accuracy of implicit tooth-tooth or teeth-gingiva boundaries [18]. Multiple segmentations can be obtained from the same input mesh by adding perturbations, such as random rotation and translation in small scales. Afterwards, the canary module will compute a confidence score for each mesh, and invoke alert for automatic self-correction or human involvement if the score is lower than a threshold value, which is determined by a real-world hold-out test. The entire system is composed in an end-to-end manner to support real-time inference for clinical applications.

Our model is trained with 4,000 IOS mesh data manually labeled by human experts with evenly distributed maxillary and mandible scans from 2,000 patients. A hold-out test on 200 mesh scans (100 patients) shows that our system achieves an average mean intersection over union (mIoU) score, per-face accuracy, average-area accuracy, and area under the receiver operating characteristic curve (AUC) of 93.92%, 96.94%, 98.26%, and 0.9991, respectively, which significantly outperform the state-of-the-art baselines. Besides, the inference speed is two orders of magnitude faster than human experts. Furthermore, we present a study on the clinical applicability of our model. In particular, a real-world large-scale study (1,000 scans from 500 patients) shows that 96.9% of the segmentations are satisfactory for clinical applications such as orthodontics, 2.9% automatically trigger alerts for human improvement, and only 0.2% need rework. DC-Net with such a low acceptable rework can substantially reduce the amount of human labor for manual inspection and post-correction. Overall, the comprehensive experiments corroborate that DC-Net can achieve much higher accuracies than state-of-the-art baselines, is much faster than human experts, and can largely reduce human efforts in practice. To the best of our knowledge, our system is the first deep learning based system for fully-automated teeth segmentation with its utility verified via a large-scale clinical test. In addition, an engineered version of DC-Net has been integrated into a clinical orthodontic system, amply demonstrating that deep learning has a great value in digital dental solutions.

RESULTS

Experimental Setup. We evaluate the performance of our system with two comprehensive experiments. The first experiment measures the accuracy and efficiency of our system with 200 mesh scans (100 patients), and compares them with the state-of-the-art baseline. The second one, a large-scale study with 2,000 scans (1,000 patients), evaluates the clinical and industrial applicability of our system. Both datasets are randomly collected from different hospitals in China in 2019, without overlap between them. For a fair comparison, all the mesh scans for maxilla or mandible are standardized to have 150,000 faces and denoted as F = {f_1, f_2 … f_N}, where N =150,000 is the number of faces. Every face fi is associated with a human experts' annotation y_i, where $y_i \in \{0, 11-18, 21-28, 31-38, 41-48\}$ denotes the FDI notation for gingiva and 32 different teeth.

Baseline and Evaluation Metrics. We select Xu's work as our baseline since it outperforms other deep learning models and traditional geometry-based methods [11-13]. We use three metrics to evaluate the segmentation results, i.e., the mean intersection over union (mIoU), the per-face accuracy and the average-area accuracy. The mIoU is the most popular metric to measure the performance of segmentation models. The per-face accuracy measures the number of faces that are correctly classified and is defined as $ACC_f = \frac{1}{N}\sum_{i=1}^{N} \Phi(y_i, \hat{y_i})$, in which $\hat{y_i}$ is the predicted label for face f_i and $\Phi(\cdot)$ is the indicator function, whereas $\Phi(y_i, \hat{y_i}) = 1$ if $y_i = \hat{y_i}$, otherwise, $\Phi(y_i, \hat{y_i}) = 0$. The average-area accuracy measures the total area of the faces that are correctly classified and is defined as $ACC_a = \sum_{i=1}^{N} \Phi(y_i, \hat{y_i}) \cdot a_i / \sum_{i=1}^{N} a_i$, where $a_i$is the area for face f_i.

Segmentation Performance. The segmentation results are reported in Table 1. For simplicity, we use DL to denote the deep learning segmentation component in Figure 1, and DLBS to denote DL and the boundary smoothing component. We can see that a single DL component can outperform the current state-of-the-art baseline in terms of all three metrics. With the boundary smoothing part to further refine the segmentations on tooth-tooth and teeth-gingiva boundaries, DLBS can achieve per-face and average-area accuracies of 96.90% and 98.17% for the mandible, and 96.98% and 98.34% for the maxillary, respectively, surpassing the baseline by 2.54-5.16%. The improvement regarding the mIoU score, which is usually considered as a better metric than accuracies to evaluate semantic segmentation models, is even more substantial, i.e., 8.29% for mandible and 4.55% for maxillary. Furthermore, we can notice that while the performance of the baseline model varies materially across the mandible and maxillary, our model exhibits consistent results on both of them. Consequently, our model is much more appealing to be integrated into current clinical solutions.

Model Complexity & Clinical Utility. To assess the model complexity and clinical utility of our model, we compare its number of parameters and inference time to those of the baseline and human experts. The results are reported in Table 2. It can be seen that our model has four times as many parameters as the baseline, due to the use of more powerful Edge-Conv blocks and the deeper architecture. Nonetheless, the end-to-end inference time, which is measured as the time for the neural networks and smoothing algorithms to generate segmentation outputs for input mesh scans, of our deep learning module is one order of magnitude faster than the baseline. This is because the baseline model relies on a plenty of predefined features of mesh faces, which would take about 5 mins to compute in practice. In contrast, our model directly operates on point clouds, which can be sampled from the input mesh on the fly. Moreover, our model is two orders of magnitude faster than human experts. Albeit one may assume human experts can always give clinically applicable segmentations, their inefficiency is often a crucial bottleneck in practical applications. This also constitutes the main motivation for our proposal of the fully-automated and fault-aware DC-Net system.

Statistical Analysis. We conduct a comprehensive statistical analysis to characterize the performance of DL over each individual tooth and the gingiva. The results are reported in Table 3 and 4. Results for the DLBS and the baseline model are attached in the Supplementary as we could not compute the AUCs for them because they only output classified categorical labels but not probabilities. We compute the multi-class AUCs with a one-vs-all strategy. The AUCs for all the teeth and the gingiva are above 0.99, with an average AUC of 0.9991 and 0.9992 for maxillary and mandible, respectively. The gingiva and few teeth, such as the left maxillary and right mandible third molars (tooth 28 and 48), perform slightly worse than others as illustrated in Figure 1.

While the mean sensitivity, specificity, positive-predicted values (PPV), and negative predicted values(NPV) of DL are slightly lower than DLBS, both of them attain significant gain over the baseline. For instance, both DL and DLBS can achieve a mean specificity of 92.72-94.36%, while the baseline can only obtain mean specificities of 84.07% and 87.92% (cf. Supplementary Table. 1-4). The DL and DLBS also universally outperform the baseline for individual teeth and the gingiva. As for the individual teeth, an easily observable finding is that the specificities of DL for the third molars in both the mandible (68.20%, 68.40%) and maxillary (80.45%, 84,18%) are much worse than that of all other teeth, though they are much higher than that of the baseline. This is mainly because that many patients do not have the third molars, as suggested by the smaller values of n* in Table 3 and 4. Hence, our model is not effectively trained to recognize them. Besides, the third molars are tightly fused the noisy gingiva part in the edge of mesh scans, and some are even not correctly scanned by 3D mouth scanners, leading to the reported inferior performance as well.

The confusion matrices illustrate the discordance between DLBS's predictions and the golden standard by human experts (Figure 3). The two confusion matrices exhibit a similar pattern and are consistent with results in Table 3 and 4, principally demonstrating that our model can correctly recognize the gingiva and individual teeth while making minor mistakes for the four third molars.

Large-Scale Clinical Applicability Study. In this work, we conduct the first large-scale study to verify the clinically applicability of deep learning solutions for 3D teeth segmentation. Specifically, a total of 1,000 mesh scans from 500 randomly selected patients with severe malocclusion are examined by DC-Net following the pipeline in Figure 1. Upon receiving a segmentation by the deep learning module, instead of calling for human improvement, the canary module will immediately decide whether it is clinically applicable or problematic based on an estimated confidence score. Meanwhile, a committee of human experts will assess the clinical applicability of all the 1,000 segmentations and their consensus would be regarded as the golden standard. The confusion matrix for this study is displayed in Table 5. We notice that 94% of the segmentations are deemed to be clinically applicable by human experts. For the 60 problematic cases, 29 of them could be automatically corrected by repeating the segmentation pipeline in DC-

Net with random data augmentations to the input point clouds. Consequently, only 31 out of the 1,000 scans are actually problematic for clinical use. Moreover, DC-Net would automatically invoke alert for human participation for 29 out of the 31 scans, leaving only 0.2% rework rate in practice. Our industrial and clinical partners confirm that such a performance is incontrovertibly acceptable for many industrial and clinical applications, such as assisting dentists to explain specific oral disease to patients and to design treatment plans for orthodontics or dental implant.

Visualization. We visualize three segmentations in Figure 4 to illustrate the major limitations of the baseline for complicated cases. More randomly chosen visualizations in Supplementary further shows some mistakes made by the baseline and demonstrate the superiority of our system. The first case shows that the baseline model fails for crowded teeth. It wrongly recognizes two incisors as one, recognizes the canine teeth and first premolar as the canine teeth, and consequently, mislabeled all the remaining teeth as shown in the right half jaw. In the second case, the baseline fails to identify the small unerupted teeth by mislabelling them as the gingiva. A mistake like this would still achieve a very high accuracy, but such segmentations are clinically infeasible. The third case illustrates two more general mistakes in most teeth segmentation systems, which are failing to identify the teeth-gingiva boundary and the third molars. Our segmentation system showed much better performance in these cases, as verified in the large-scale study, with a very low frequency of mislabeling. In the Supplementary, we can further notice that the baseline is error-prone with its hierarchical segmentation scheme while our system maintains consistent performance. For example, the baseline misclassifies a large part of teeth or gingiva in its first teeth-gingiva segmentation stage in some extreme cases, as shown in Figure S1-S3. In contrast, our system never commits such mistakes. These mistakes could not be fixed in the tooth-tooth segmentation stage and would cause severe repercussion in practice.

DISCUSSION

An efficient and accurate 3D teeth segmentation system is indispensable in digital dentistry. It could help dentists to identify oral disease more accurately, communicate with patients more clearly, and design and simulate treatment plans more efficiently. In this paper, we propose the first fully-automated and clinically applicable 3D teeth segmentation system by deep learning, which significantly outperforms previous methods and is readily integrated in a real-world clinical orthodontics system. Other promising potential applications of this system include dental implant, restoration, and complete denture.

One of the key novelties in our system is that we solve the original mesh segmentation task via segmentation on point clouds with an end-to-end deep neural network. Zanjani et al. exhibited a similar strategy by using PointCNN for 3D teeth segmentation. However, their performance is worse than the baseline and our system, and the model was not comprehensively evaluated in terms of clinical utility and applicability. Some studies argued that point clouds can only capture coarse local dependencies while ignoring information from mesh surfaces, and as a result, the related deep learning methods are not suitable for dense classification tasks such as fine-grained teeth segmentation [19,20]. However, our experiments reveal that, by utilizing both the point positions and geometrical information such as face normals and shapes, deep neural networks could achieve better performance on point clouds than on mesh surfaces which contain richer geometrical information [11,19,20]. Moreover, sampling such inputs from a mesh scan can be done on the fly, and the subsequent inference step only costs 4 seconds with a low-end NVIDIA

1050Ti GPU. Whereas, methods working on meshes usually suffered from a lower efficiency or inferior generalization ability. For example, Xu et al. spent at least 5 minutes to extract 600 predefined features from mesh faces [11]. The model proposed by Lian et al. was only trained to handle scans with regular morphology and without the third molars, hence producing problematic segmentation for complicated cases, such as scans with missing teeth or third molars [19,20]. In addition, though they operated on mesh data, a graph-cut smoothing algorithm was still required for boundary refinement.

Some studies followed a hierarchical procedure to generate segmentation outputs. For example, Xu et al. first distinguished teeth from gingiva and subsequently classified each tooth. Tian et al. synthesized gingiva, incisors, molars and premolars step-by-step. These strategies are time-consuming and error-prone, as mistakes occurred in former steps could not be corrected in later steps. Similar to other research, our deep neural network performs segmentation in an end-to-end manner to improve clinical utility.

In addition, different deep learning architectures possess distinctive representation learning capabilities and hence varied accuracies. Xu et al. and Tian et al. emphasized more on feature extraction. Their proposed neural networks, based on conventional 2D or 3D CNNs, are not specified for mesh segmentation tasks. The network in Zanjani et al. only considers the 3D coordinates information of each point, ignoring the geometrical information such as face normals or shapes and the categorical jaw information. Sun et al. and Lian et al. designed specific networks to learn from meshes or vertices, but the robustness and generalization ability of their models are yet unsatisfied or not verified for clinical applications. In our study, we sample rich geometrical information from mesh data during pre-processing. Besides, we design the deep learning architecture with the EdgeConv block, which could capture the geometric relationship between each point and its neighbors with an explicit local graph constructed by the kNN algorithm. We further stack multiple EdgeConv blocks and fuse representations from different levels to learn a semantically expressive global representation for segmentation. Consequently, our system performs better for complicated dental morphology specificities.

Another marked contribution of this study is that we carried out a series of experiments with real-world clinical data and conducted the first large-scale clinical applicability test for the proposed segmentation system, which demonstrates that deep learning has a great potential in digital dental solutions. The reported performance of previous methods were only tested with a small amount of patient data, which would degrade a lot for real-world clinical data with complicated dental morphologies. For example, previous research reported a labeling accuracy above 98% in their experiment [11], however, its accuracy drops to 91-94% with real-world clinical data and is not consistent across the maxillary and mandible as displayed in Table 1. In stark contrast, our system achieves consistent and significantly better performance, especially for the widely used mIoU metric. Furthermore, the large-scale study with an industrial company and its clinical partners verified the industrial and clinical applicability of our system, which alternatively reveals that our system could help to largely reduce human efforts for manual inspection and post-processing in teeth segmentation.

Furthermore, reliable and interpretable decision making is crucial for clinical applications. Previous methods did not provide any confidence estimations for the decision making process. Along with the degraded performance, they could not be directly integrated into clinical or industrial applications as numerous human interventions are still required to ensure reliable diagnosis. This limitation is addressed by the canary module in our system. The canary module could compute an estimated confidence score for each mesh scan (see Methods) and decide whether the segmentation is clinically applicable or not. The threshold value used to determine the clinical applicability is chosen by a committee of human experts to ensure both the rework rate and false negative rate are acceptable in clinical applications.

There are nevertheless some limitations in this study for future work. First, our system still generates few suboptimal segmentations, i.e., 31 out of 1,000 scans need human improvement. Most of the failures come from very complicated dental morphologies, such as supernumerary teeth or extremely crowded teeth. Besides, the performance regarding the third molars need to be improved as well. Acquiring more related data or designing novel neural networks might help mitigate these issues. Second, although our system is composed in an end-to-end manner, the DL and BS parts are still separated. The BS costs 20 seconds on average during inference, which is 5 times more than DL. A better solution is to design a deep neural network which can perform BS by itself or does not require boundary smoothing at all. This would lead to a much more efficient segmentation system. Finally, the integration of our system to other digital dental applications besides orthodontics, such as dental implant or complete denture, should be independently verified with large-scale clinical or industrial tests.

CONCLUSION

In this study, we proposed an accurate, efficient, fully-automated, and most importantly, clinically applicable 3D teeth segmentation system. Our system significantly outperforms the state-of-the-art methods and has been integrated into real-world clinical orthodontic systems to serve patients and dentists. The large-scale clinical applicability study demonstrates that deep learning is a potent element to intelligent digital dentistry. Prospective research combining artificial intelligence and dental medicine would make a big difference in future digital dentistry.

METHODS

DC-Net Overview. The DC-Net includes a segmentation module and a canary module. The segmentation module includes the data preprocessing, deep learning based segmentation, and boundary smoothing components. The canary module includes the confidence evaluation and auto-correction components. The system is illustrated in Figure 1. Below are the details for each component.

Data Preprocessing. The data preprocessing component takes the original IOS 3D teeth mesh scans as the input, automatically aligns the meshes, and generates the corresponding point clouds with a predefined resolution. As shown in Figure 1, the original IOS data is first collected via 3D scanners in clinics, which is of high resolution with more than 150,000 mesh faces. In practice, different clinics might use different laser scanners, resulting in large variations of positions in a 3D Cartesian space for the 3D IOS mesh scans. To obtain mesh data with standard reference positions, we leverage on the widely-used Iterative Closest Point (ICP) [13] registration algorithm provided by Open3D to align all mesh data to predefined positions [19]. More specifically, we establish three mesh templates based on the dental arch shapes, which determine the major teeth morphologies. For each input mesh, three alignment scores are computed based on the overlapping areas, which quantitatively measures the inliner correspondence between the input mesh and the templates. The alignment with the highest score is selected for subsequent steps.

We transform the aligned mesh to a point cloud and extract predefined features. The motivation to use point clouds is as follows. First, the IOS mesh is of very high resolution with more than 100,000 faces, and training a deep neural network with such inputs might cause the GPUs to be overloaded. Second, compared to the complicated formality of mesh data, point clouds are much easier to process with deep neural networks. Finally, there are plenty of research works in deep learning for point clouds representation learning while much less for meshes. To obtain the point clouds, we first uniformly sample 10,000 faces from the original mesh and regard each face center as a point to form a point cloud with 10,000 points. We did not use all the 100,000+ mesh face centers due to efficiency reasons. The features associated with each point include three parts: 1. the position of the face, i.e., the 3D coordinate of the face center denoted as $h_c = (x_c, y_c, z_c) \in R^3$; 2. the 3-dimensional normal vector of the face surface $h_n \in R^3$; 3. the face shape feature $f_s \in R^9$. For each face with three vertices as $(x_i, y_i, z_i)$for i from 1 to 3, and a face center (x_c, y_c, z_c), the face shape feature is defined as $Concat(x_i - x_c, y_i - y_c, z_i - z_c)$ for i from 1 to 3. $Concat(\cdot)$ is the concatenate operation for vectors, leading to a 9-dimensional shape feature for each face. In consequence, the final output after data preprocessing for each IOS mesh scan is a point cloud with 10,000 points, each associated with a 15-dimensional feature vector $h = Concat(h_c, h_n, h_s) \in R^{15}$. This preprocessing procedure can be finished with a modern computer on-the-fly.

Deep Learning Based Segmentation. The goal of the deep learning based segmentation system is to perform point-wise classification for the point clouds obtained in the data preprocessing step. Following the FDI standard, we define 33 labels within the point clouds, where $y_i \in \{ 11 - 18, 21 - 28\}$are used for the maxilla, $y_i \in \{31 - 38, 41 - 48\}$are used for the mandible and $y_i = 0$denotes the gingiva. Formally, given an F-dimensional point cloud with n points, denoted as $P = \{p_1, p_2, \dots p_n\} \in R^F$where F=15 with the preprocessing step, our goal is to assign a label $\hat{y_i}$ for each point p_i, where $\hat{y_i} \in \{ 0, 11 - 18, 21 - 28, 31 - 38, 41 - 48\}$.

Though point clouds provide a more flexible representation for 3D objects than mesh, there are still some specific concerns when processing them with deep neural networks. A critical issue is that point clouds are unordered and lack topological information. Hence, to maintain consistency over segmentation results, the deep neural network needs to be invariant to the permutations. As for teeth segmentation, capturing the topological information is critical to deal with complicated morphology specificities, such as crowded teeth, shape deformation and dentitional abnormality etc, which usually require rich geometrical information. The local topological structure is also beneficial to identify the boundary points among teeth or between teeth and the gingiva, which can help with complicated cases such as gingival recession. Previous methods usually fail to precisely recognize such complicated morphological specificities. Addressing the aforementioned issues requires a more expressive and powerful system.

The architecture of the deep learning based segmentation system is illustrated in Figure 5. This architecture is inspired by Dynamic Graph CNN (DGCNN) [15] with modifications to adapt to the 3D teeth data which is of much higher resolution and morphological complexity. In our system, the input point cloud is first transformed to a canonical space by a 3x3 matrix which is estimated with the coordinates of each point and its k neighbors by the Transformation Net [16]. After that, the transformed input is fed into the Edge-Conv block [15], which is the key building block in our architecture. In DC-Net, an Edge-Conv block is composed of the kNN feature extractor and three 2D convolutional layers with permutation invariant aggregation operations. The Edge-Conv block aims to capture topological structure by constructing an explicit local graph with edges among neighborhood points and updating the edge features with convolutional operations. The edge feature can help capture the geometrical relationship for each point and its neighbors. The neighbors of a given point are computed with the k-nearest neighbor(kNN) algorithm based on the latent representations. Hence, the proximity is changed from layer to layer, and the corresponding neighborhood graph is dynamically updated as well. By stacking multiple Edge-Conv blocks, the neural network is able to learn both local geometric features in bottom Edge-Conv blocks and non-local semantic relationships in top Edge-Conv blocks. The detailed settings of each Edge-Conv block are described in the Supplementary.

We concatenate both the mean and max pooled outputs for the first two Edge-Conv blocks to preserve richer information to learn meaningful global representations. A categorical vector which describes whether the scan is for maxillary or mandible is also encoded with a convolutional layer and then fed into the network. It could be regarded as global prior information to the system. In experiments, we found that our network never assigns maxillary labels to mandible inputs or vise versa. Local representations from different Edge-Conv blocks are concatenated together to be processed by a series of convolutional layers to reach a global representation. We use dropout with a dropout ratio of 0.6 to improve the generalization ability of our network .

We apply data augmentation with four different perturbations to enlarge the training set and improve the generalization ability of DL over spatial transformations. We independently sample 10,000 points every time for each perturbation. The perturbation is either random rotation by -10 to 10 degrees, or random translation by -10 to 10 millimeters, along one of the x, y, and z axes. The network is trained for 100 epochs. The model is evaluated on a validation set with 100 scans and the model with the highest validation accuracy is selected for testing. During testing, we sample 30,000 points for each mesh, which will be segmented by DL. For the rest points (i.e., faces), we choose five coordinate-based nearest neighbors from the 30,000 points, and the neighbor with the highest probability will determine the label for that point. Neither the per face accuracy nor average area accuracy is sensitive to the number of neighbors from 5 to 50. Another strategy is to use DL to segment all the 150,000 points with a lower inference speed to avoid neighbor selection. Both inference methods give consistent results.

Boundary Smoothing. The segmentation result produced by DL might be coarse around the tooth-tooth and teeth-gingival boundaries. We use the multi-label graph-cut based optimization strategy for boundary smoothing to further improve the performance[18].

Let F be the set of faces on the mesh, f_i be the i-th face, and f_j be the adjacent faces of f_i . We use y_i as the label assigned to f_i with the probability q_i generated by DL where $y_i \in \{ 0, 11-18, 21-28, 31-38, 41-48\}$. With the graph-cut algorithm, the labels y_i are determined by optimizing over the objective function as

$$min \quad \sum_{f_i \in F} \quad E_1(y_i) + \lambda \sum_{f_i, f_j \in F} \quad E_2(y_i, y_j) \ ,$$

where $E_1(y_i) \ = \ - log \ (q_i)$ and $E_2(y_i, y_j) = - \ log \frac{\theta_{ij}}{\pi} \phi_{ij} \ if \ y_i \ \neq y_j; \ otherwise \ E_2(y_i, y_j) = 0.$

Intuitively, $E_1$is the log-likelihood for label assignment. If we assign $y_i$which is associated with a lower probability q_i to f_i, it would incur a large penalty in the objective function. $\theta_{ij}$is the dihedral angle between f_i and f_j, and ϕ_{ij} is the distance between the face centers of f_i and f_j. E_2 reinforces the prior that adjacent faces shall share consistent labels. λ is a configurable nonnegative smoothing penalty to balance the two terms. As indicated in Table 1, applying boundary smoothing to the segmentations by DL will improve the accracies a bit.

Confidence Evaluation and Auto-correction. We develop the canary module with the confidence evaluation and auto-correction components to build a more reliable, interpretable and fully-automated system. The goal of canary is to catch segmentation failures as much as possible while maintaining a low false alarm rate. This is equivalent to a near-zero false negative rate and a low false positive rate. We estimate the confidence score of the segmentations by quantitatively measuring its robustness against 6 different small spatial transformations, as we notice that most of the segmentation failures occurred on meshes where small perturbations can lead to substantially inconsistent segmentations. The spatial transformations are denoted as $T(translation, rotation) = \{(5,0), (-5,0), (0, \frac{\pi}{18}), (0, \frac{-\pi}{18}), (0, \frac{\pi}{36}), (0, \frac{-\pi}{36})\}$, where the translation and rotation are defined with a millimeter and radius scale, respectively. For simplicity, we denote the segmentation for the original mesh by DLBS as S_0, and segmentations for the 6 perturbed meshes by DL as $S_1\ to\ S_6$. For efficiency reasons, we do not smooth the 6 perturbed segmentations for confidence evaluation.

We compute the confidence score C based on the similarity between $S_0$and $S_1\ to\ S_6$, i.e, $C = min\{sim(S_0, S_i)\}, for\ i = 1\ to\ 6$, where $sim(\cdot)$ measures the similarity between two segmentations. Let's take $sim(S_0, S_1)$ as an example to illustrate the details. We first define a similarity score for each of the 33 classes in segmentation, which is computed as $sim_l = min\ (\frac{\sum_{n=1}^{N} I(f_n{}^0 = f_n{}^1 = l\,)}{\sum_{n=1}^{N} I(f_n{}^0 = f_n{}^1 = l) + \sum_{n=1}^{N} I(f_n{}^0 = l\,, f_n{}^1 \neq l)}, \frac{\sum_{n=1}^{N} I(f_n{}^0 = f_n{}^1 = l\,)}{\sum_{k=1}^{K} I(f_n{}^0 = f_n{}^1 = l) + \sum_{n=1}^{N} I(f_n{}^0 \neq l\,, f_n{}^1 = l)})$, where N is the number of faces, l is the corresponding label, and $f_n{}^0$ and $f_n{}^1$denote the labels for the n-th face in S_0 and S_1, respectively. $I(\cdot)$ is an indicator function with multiple conditions. Afterwards, we define the similarity score between S_0 and S_1 as $sim(S_0, S_1) = min\ _{l \in L}\{sim_l\}$, where $L = \{\,0, 11-18, 21-28, 31-38, 41-48\}$. This score, which is conceptually similar to the IoU, measures the consistency and robustness of the segmentations. Finally, we will determine whether S_0 is clinically applicable by comparing $C$to a configurable alert threshold value ϵ. The experiment with a committee of human experts over a hold-out set with 200 scans indicates that $\epsilon = 0.85$ can give a zero false negative rate and a 10% false positive rate, which are acceptable for clinical and industrial applications. Hence, we use $\epsilon = 0.85$ for our clinical applicability test in this study.

If a certain input triggers alerts, we will smooth all the other 6 segmentation results used in confidence evaluation for auto-correction. The one that produces the minimal objective in boundary smoothing becomes the final output. Since during training, we augment the dataset by randomly perturbing the mesh in a small scale for 5 times, it has a fair chance that DLBS can produce better segmentation results on perturbed mesh over the original one. The results in the large-scale study confirmed this assumption.

Statistical Analysis. The mathematical formulation for segmentation metrics, such as mIoU, per-face accuracy and average-area accuracy, are defined in the main paper and Supplementary. The AUC is computed with a one-vs-all strategy to evaluate the performance of DL. The sensitivity, specificity, positive predicted values (PPV) and negative predicted values (NPV) are also used to evaluate the performance. All statistical analysis is conducted with Python and the corresponding packages such as sklearn and scipy.

Reproducibility. The model and its performance are verified with independent test cohorts. We also verified the performance of DL with multiple random seeds or initializations or Tensorflow versions or on different hardware platforms.

Data Availability. More visualization results are available in Supplementary. We may offer a limited number of data samples for academic use upon request. The whole dataset is not publicly available due to privacy policies.

Code Availability. The implementation code could be shared for academic use upon request.

ACKNOWLEDGEMENTS

We acknowledge helpful discussions with K.Chen, H.M.Li, H.K.Wu, and H.X.Qiu.

AUTHOR CONTRIBUTIONS

J.H., Z.Z.L., and Z.H.Z. designed and initiated the project. J.P., Z.Z., Z.C., B.W.Z., and Y.F. designed the model architecture. W.L., Y.L.Z., P.L.L., J.R.Y., W.X.S., and T.Y.C. conducted validation experiments. Y.F., H.H.Y., H.J.H., E.P.L. and B.F. created the dataset. J.H. and Z.Z.L. wrote the manuscript. J.P. and Z.Z. developed the software.

DECLARATION OF INTERESTS

J.P., Z.Z., and Z.C. are employed in DeepAlign. B.W.Z. and Y.F. are employed in Angel Align. J.H. and Z.Z.L. are consultants to DeepAlign. None of the other authors declare competing interests.

FIGURE LEGENDS

Figure 1: System Design of DC-Net. The DC-Net system is composed of two main modules to generate clinically applicable segmentation outputs. The deep learning segmentation module includes data preprocessing, deep learning segmentation and boundary smoothing. The canary module includes confidence evaluation, auto-correction, and fault-alarm for potential human improvement.

Figure 2: ROC curves for the DL model. ROC curves for (A): ROC curves for the gingiva in maxillary and mandible. (B-C): ROC curves for three teeth in mandible and maxillary, respectively. We selected the worst-performed teeth (teeth 18 and 48); 2 randomly chosen canine teeth (teeth 11 and 32) and 2 randomly chosen molar teeth (teeth 25 and 36). The horizontal axis is zoomed for better visualization in (A-B).

Figure 3: Confusion Matrices. The confusion matrix for the predictions by DLBS versus the ground-truth from human experts for the mandible (A) and maxillary (B), respectively. The percentage of all mesh faces in each category is displayed on a color gradient scale.

Figure 4: Segmentation Results Comparison between the Baseline and Our System. Details of the selected boxes are displayed in the corresponding zoom-in boxes on the right. The teeth and gingiva are labeled with 17 different predefined colors, with a reference shown in the Supplementary.

Figure 5: Deep Neural Network Architecture. The deep learning architecture takes the point clouds as inputs and outputs the segmentation results which could be mapped back to meshes easily using kNN. The network first transforms the set of points to a canonical space with the Transformation Net. Then, the EdgeConv blocks will compute edge features for each point, and aggregate features using 2D convolutional layers. We stack both mean- and max-pooling layers after the first two EdgeConv blocks, and the pooled outputs will be concatenated as the input for the next block. The outputs from all EdgeConv blocks are concatenated together to form a global feature descriptor before the one-hot encoded categorical vector (maxilla or mandible) is fed into the network. Finally, we stack four 2D convolutional layers, which aggregate the concatenation of the outputs from all intermediate EdgeConv blocks and the global feature descriptor, to generate point-wise classification scores for 33 semantic labels. $\oplus$ in the diagram stands for concatenation. The detailed settings of the architecture are listed in Methods and Supplementary.

REFERENCES

1. Yuan, T., Liao, W., Dai, N., Cheng, X. & Yu, Q. Single-Tooth Modeling for 3D Dental Model. *Int J Biomed Imaging* **2010**(2010).
2. Herrmann, W. [On the completion of Federation Dentaire Internationale specifications]. *Zahnarztl Mitt* **57**, 1147-1149 (1967).
3. Fan, R., Jin, X. & Wang, C.C. Multiregion segmentation based on compact shape prior. *IEEE Transactions on Automation Science and Engineering* **12**, 1047-1058 (2014).
4. Kondo, T., Ong, S.H. & Foong, K.W. Tooth segmentation of dental study models using range images. *IEEE Transactions on medical imaging* **23**, 350-362 (2004).
5. Kumar, Y., Janardan, R., Larson, B. & Moon, J. Improved segmentation of teeth in dental models. *Computer-Aided Design and Applications* **8**, 211-224 (2011).
6. Li, Z., Ning, X. & Wang, Z. A fast segmentation method for STL teeth model. in *2007 IEEE/ICME International Conference on Complex Medical Engineering* 163-166 (IEEE, 2007).
7. Wu, K., Chen, L., Li, J. & Zhou, Y. Tooth segmentation on dental meshes using morphologic skeleton. *Computers & Graphics* **38**, 199-211 (2014).
8. Zou, B.-j., Liu, S.-j., Liao, S.-h., Ding, X. & Liang, Y. Interactive tooth partition of dental mesh base on tooth-target harmonic field. *Computers in biology and medicine* **56**, 132-144 (2015).
9. Zhao, M., Ma, L., Tan, W. & Nie, D. Interactive tooth segmentation of dental models. in *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference* 654-657 (IEEE, 2006).
10. Li, Z. & Wang, H. Interactive tooth separation from dental model using segmentation field. *PloS one* **11**, e0161159 (2016).
11. Xu, X., Liu, C. & Zheng, Y. 3D tooth segmentation and labeling using deep convolutional neural networks. *IEEE transactions on visualization and computer graphics* **25**, 2336-2348 (2018).
12. Tian, S.*, et al.* Automatic classification and segmentation of teeth on 3D dental model using hierarchical deep learning networks. *IEEE Access* **7**, 84817-84828 (2019).
13. Zanjani, F.G.*, 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* 557-571 (2019).
14. Lian, C.*, et al.* Deep Multi-Scale Mesh Feature Learning for Automated Labeling of Raw Dental Surfaces from 3D Intraoral Scanners. *IEEE Transactions on Medical Imaging* (2020).

15. Wang, Y.*, et al.* Dynamic graph cnn for learning on point clouds. *Acm Transactions On Graphics (tog)* **38**, 1-12 (2019).
16. Qi, C.R., Su, H., Mo, K. & Guibas, L.J. Pointnet: Deep learning on point sets for 3d classification and segmentation. in *Proceedings of the IEEE conference on computer vision and pattern recognition* 652-660 (2017).
17. Qi, C.R., Yi, L., Su, H. & Guibas, L.J. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. in *Advances in neural information processing systems* 5099-5108 (2017).
18. Boykov, Y., Veksler, O. & Zabih, R. Fast approximate energy minimization via graph cuts. *IEEE Transactions on pattern analysis and machine intelligence* **23**, 1222-1239 (2001).
19. Sun, D.*, et al.* Automatic Tooth Segmentation and Dense Correspondence of 3D Dental Model. in *International Conference on Medical Image Computing and Computer-Assisted Intervention* 703-712 (Springer, 2020).
20. Lian, C.*, et al.* MeshSNet: Deep Multi-scale Mesh Feature Learning for End-to-End Tooth Labeling on 3D Dental Surfaces. in *International Conference on Medical Image Computing and Computer-Assisted Intervention* 837-845 (Springer, 2019).

Figures

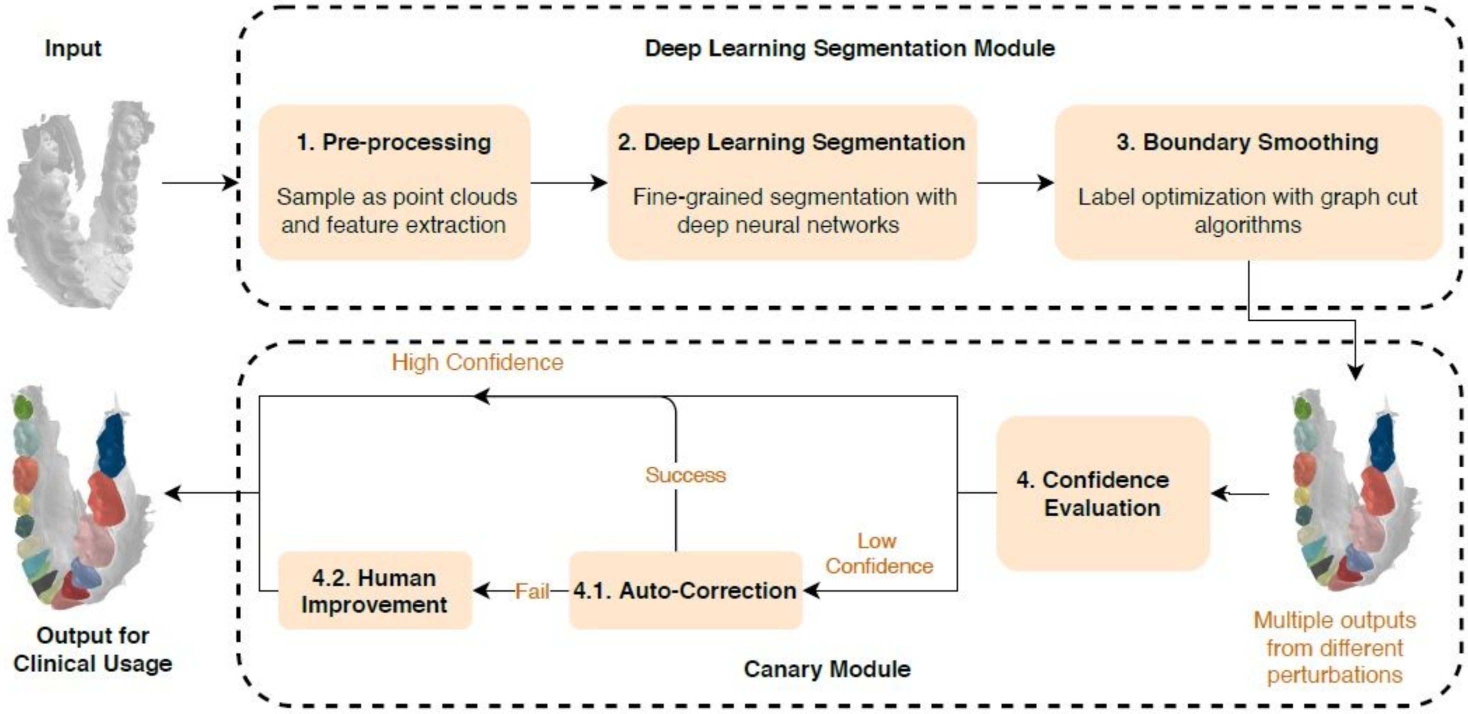
Input
Deep Learning Segmentation Module
1. Pre-processing
Sample as point clouds and feature extraction
2. Deep Learning Segmentation
Fine-grained segmentation with deep neural networks
3. Boundary Smoothing
Label optimization with graph cut algorithms
High Confidence
Success
4. Confidence Evaluation
Low Confidence
4.2. Human Improvement
Fail
4.1. Auto-Correction
Output for Clinical Usage
Canary Module
Multiple outputs from different perturbations

Figure 1

The DC-Net system is composed of two main modules to generate clinically applicable segmentation outputs. The deep learning segmentation module includes data preprocessing, deep learning segmentation and boundary smoothing. The canary module includes confidence evaluation, auto-correction, and fault-alarm for potential human improvement.

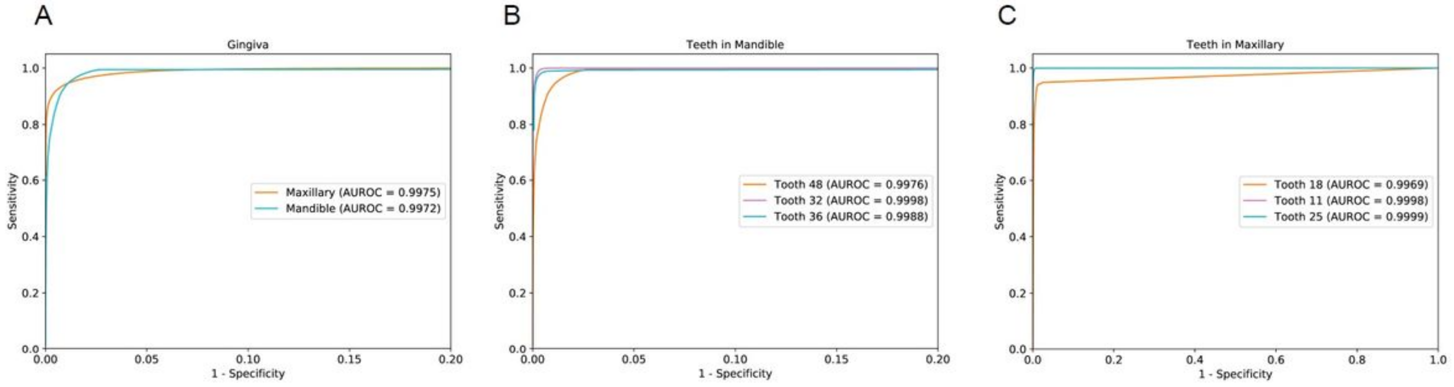

Figure 2

ROC curves for (A): ROC curves for the gingiva in maxillary and mandible. (B-C): ROC curves for three teeth in mandible and maxillary, respectively. We selected the worst-performed teeth (teeth 18 and 48); 2

randomly chosen canine teeth (teeth 11 and 32) and 2 randomly chosen molar teeth (teeth 25 and 36). The horizontal axis is zoomed for better visualization in (A-B).

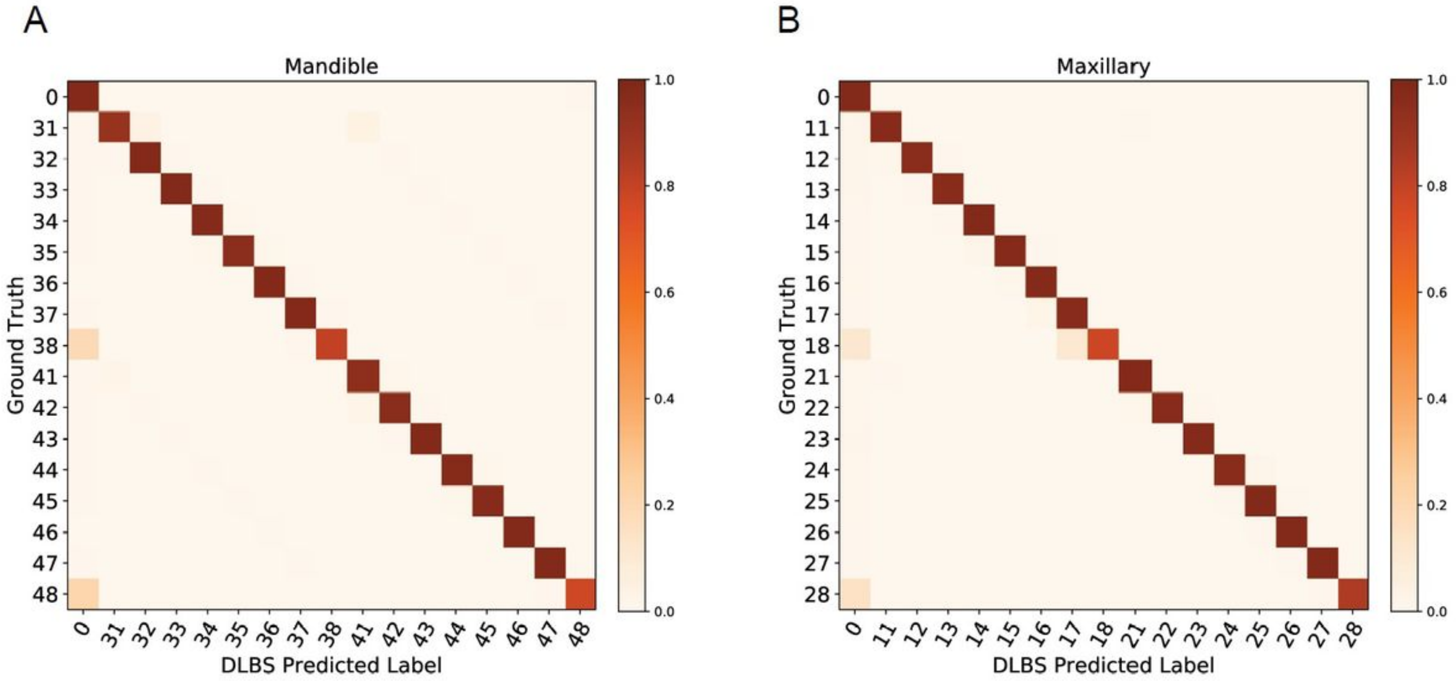

Figure 3

The confusion matrix for the predictions by DLBS versus the ground-truth from human experts for the mandible (A) and maxillary (B), respectively. The percentage of all mesh faces in each category is displayed on a color gradient scale.

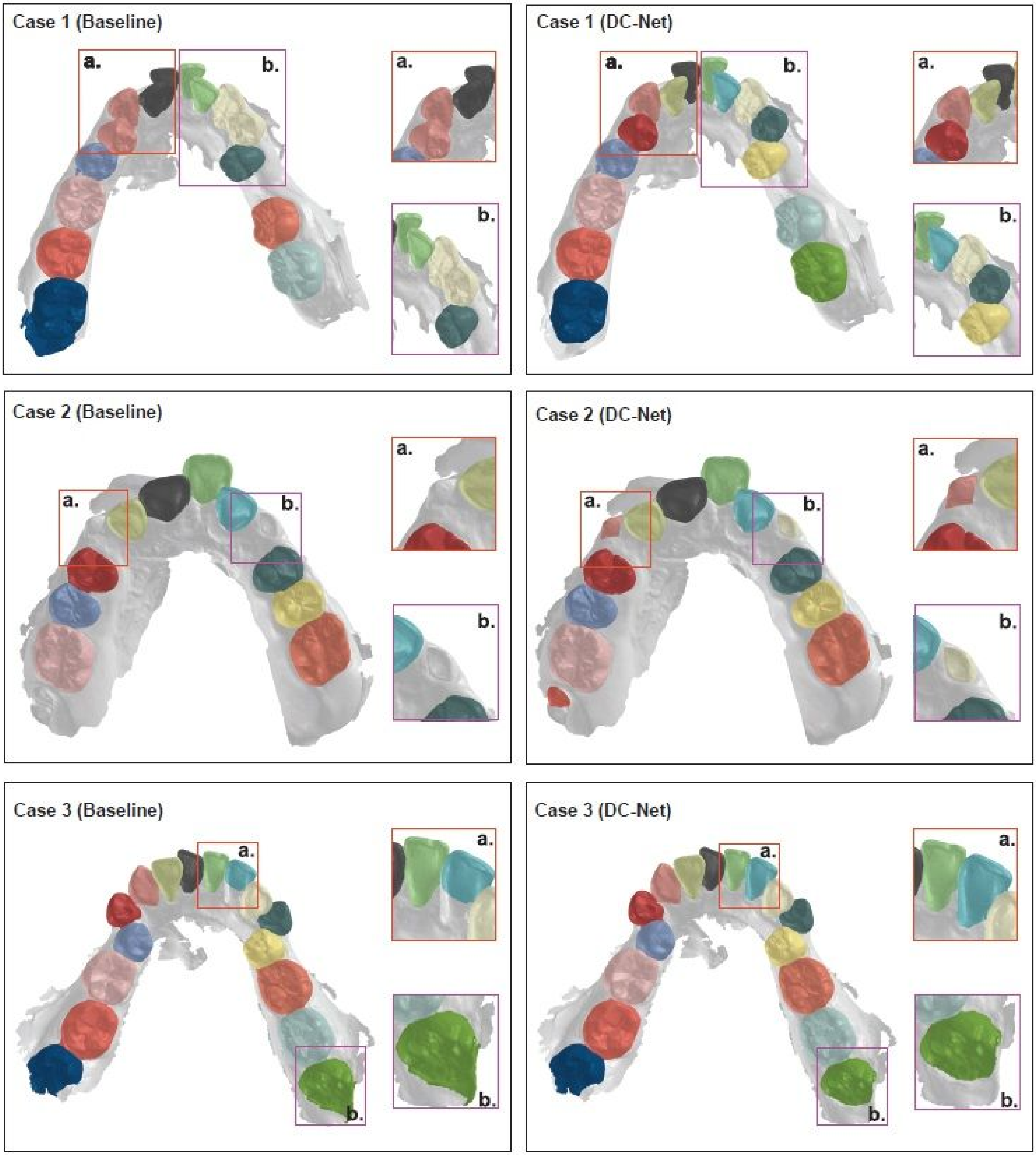

Figure 4

Details of the selected boxes are displayed in the corresponding zoom-in boxes on the right. The teeth and gingiva are labeled with 17 different predefined colors, with a reference shown in the Supplementary.

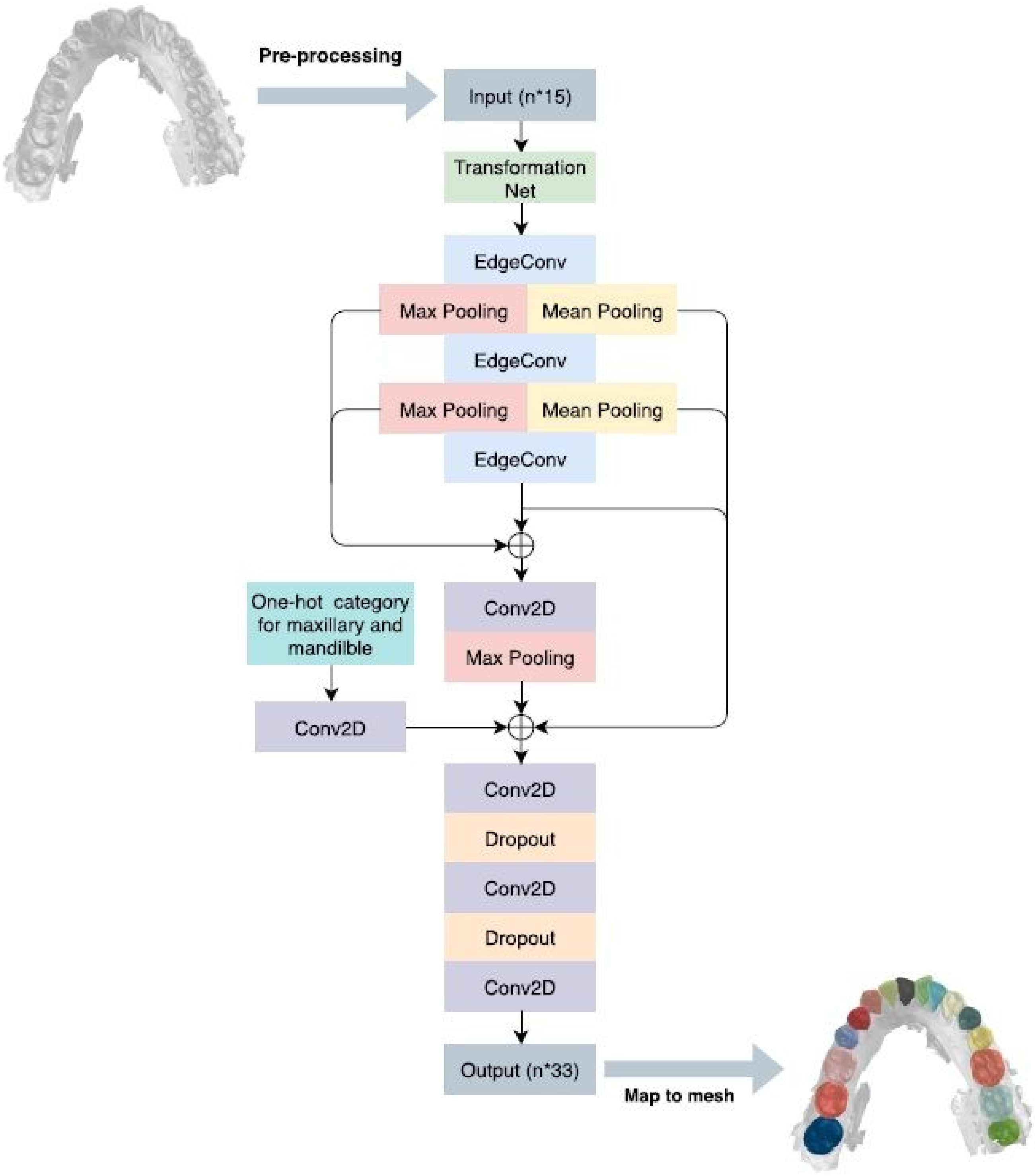

Figure 5

The deep learning architecture takes the point clouds as inputs and outputs the segmentation results which could be mapped back to meshes easily using kNN. The network first transforms the set of points to a canonical space with the Transformation Net. Then, the EdgeConv blocks will compute edge features for each point, and aggregate features using 2D convolutional layers. We stack both mean- and max-pooling layers after the first two EdgeConv blocks, and the pooled outputs will be concatenated as the input for the next block. The outputs from all EdgeConv blocks are concatenated together to form a global feature descriptor before the one-hot encoded categorical vector (maxilla or mandible) is fed into the network. Finally, we stack four 2D convolutional layers, which aggregate the concatenation of the outputs from all intermediate EdgeConv blocks and the global feature descriptor, to generate point-wise classification scores for 33 semantic labels. &oplus; in the diagram stands for concatenation. The detailed settings of the architecture are listed in Methods and Supplementary.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterials.docx
- FigureS1.pdf
- FigureS2.pdf
- FigureS3.pdf
- TABLE1.docx
- TABLE2.docx
- TABLE3.docx
- TABLE4.docx
- TABLE5.docx