Deep learning approach to semantic segmentation in 3D point cloud intra-oral scans of teeth

Proceedings of Machine Learning Research 102:557–571, 2019 MIDL 2019 – Full paper track

Deep Learning Approach to Semantic Segmentation in 3D Point Cloud Intra-oral Scans of Teeth

Farhad Ghazvinian Zanjani[1] F.GHAZVINIAN.ZANJANI@TUE.NL
David Anssari Moin[2] DAVID@PROMATON.COM
Bas Verheij[2] BAS@PROMATON.COM
Frank Claessen[2] FRANK@PROMATON.COM
Teo Cherici[2] TEO@PROMATON.COM
Tao Tan[1] T.TAN1@TUE.NL
Peter H. N. de With[1] P.H.N.DE.WITH@TUE.NL
[1] *Eindhoven University of Technology, 5600MB Eindhoven, The Netherlands*
[2] *Promaton Inc., 1076GR Amsterdam, The Netherlands*

Abstract

Accurate segmentation of data, derived from intra-oral scans (IOS), is a crucial step in a computer-aided design (CAD) system for many clinical tasks, such as implantology and orthodontics in modern dentistry. In order to reach the highest possible quality, a segmentation model may process a point cloud derived from an IOS in its highest available spatial resolution, especially for performing a valid analysis in finely detailed regions such as the curvatures in border lines between two teeth. In this paper, we propose an end-to-end deep learning framework for semantic segmentation of individual teeth as well as the gingiva from point clouds representing IOS. By introducing a non-uniform resampling technique, our proposed model is trained and deployed on the highest available spatial resolution where it learns the local fine details along with the global coarse structure of IOS. Furthermore, the point-wise cross-entropy loss for semantic segmentation of a point cloud is an ill-posed problem, since the relative geometrical structures between the instances (e.g. the teeth) are not formulated. By training a secondary simple network as a discriminator in an adversarial setting and penalizing unrealistic arrangements of assigned labels to the teeth on the dental arch, we improve the segmentation results considerably. Hence, a heavy post-processing stage for relational and dependency modeling (e.g. iterative energy minimization of a constructed graph) is not required anymore. Our experiments show that the proposed approach improves the performance of our baseline network and outperforms the state-of-the-art networks by achieving 0.94 IOU score.

Keywords: Deep learning, 3D point cloud, intra-oral scan, semantic segmentation.

1. Introduction

The emergence of digital equipment for extra-oral (e.g. X-ray panoramic, cephalometric and cone beam computed tomography) and intra-oral imaging (e.g. laser or structured light projection scanners) has been a driving force for developing computer-aided design (CAD) systems to analyze the imaging data for highly accurate treatment planning. The purpose of this paper is to explore a segmentation methodology based on deep learning for providing useful clinical information to support better treatment. For supporting an automated clinical workflow in implantology and orthodontic fields, such a CAD system should be able to resolve some fundamental issues of which accurate semantic segmentation of teeth and gingiva (gums) from imaging data is highly desirable. Here, the

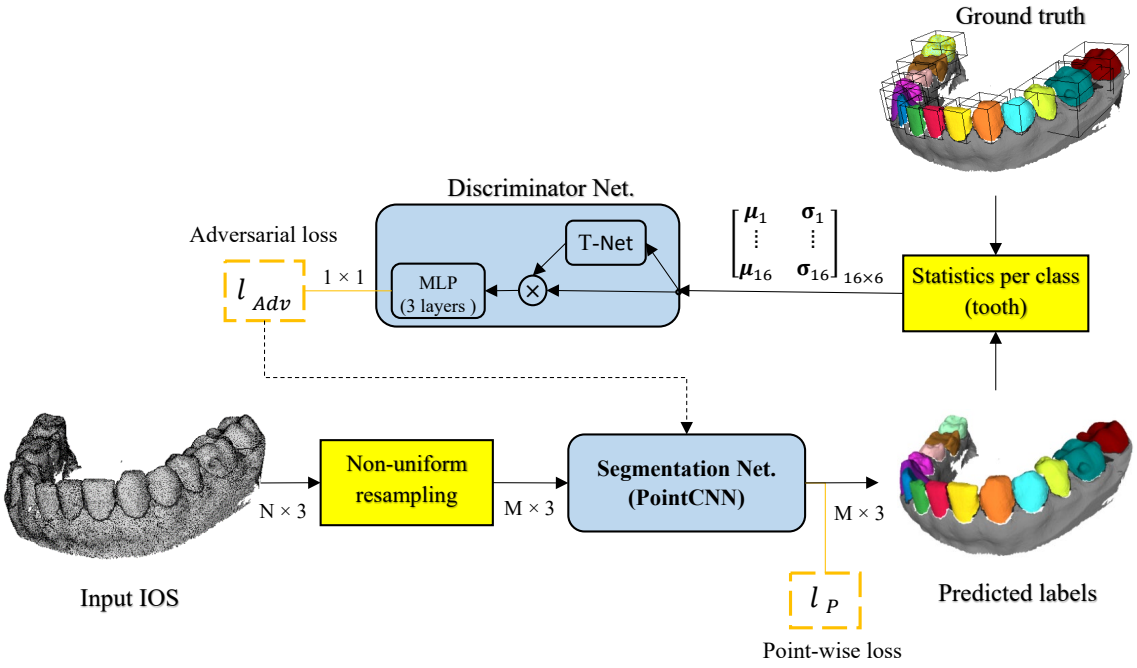

Figure 1: Block diagram of the proposed method in training mode.

semantic segmentation problem for intra-oral scans (IOS) refers to assigning a label, based on the *Federation Dentaire Internationale* (FDI) standard. In more technical details, this involves labeling all points as belonging to a specific tooth crown or as belonging to gingiva within the recorded IOS point cloud. Each point is represented by a coordinate in the 3D Cartesian coordinate system which is not universal (i.e. the latter can be different between two IOS). The FDI specifies 32 labels for adult dentition, referring to 16 teeth in each upper and lower jaw. In this study, we treat the teeth on the upper and lower jaw in the same way, so that we only employ 16 separate labels to be classified. This changes the problem to finding 16 classes (+1 extra for the gingiva), which facilitates better learning.

To bring artificial intelligence (AI) into modern dentistry, we improve IOS semantic segmentation by means of end-to-end learning of a segmentation model. Building an accurate segmentation model involves two aspects of complexity. Firstly, complexity originates from the dentition (teeth arrangement) and data acquisition. Since the shape of two adjacent tooth crowns (e.g. two molar teeth) may appear to be similar, assigning a correct label demands additional information such as relative position with respect to other teeth on the dental arch. Furthermore, presence of abnormalities in dentition and shape deformation, makes IOS segmentation a challenging task for a segmentation model. An examples of such an abnormality may be lacking teeth (e.g. wisdom teeth). Additional challenges may arise from acquisition issues such as partially missing data (e.g. because of occlusion in scanning), lack of a universal coordinate system, presence of noise, outliers, etc. The interaction of these challenges is important for successfully applying computer vision algorithms.

The second aspect of complexity relates to the 3D geometrical representation of data by a point cloud that is not well suited to the decent deep learning models that are highly performant

on 2D/3D images. Application of such deep learning models (e.g. CNN-based architectures) to point cloud analysis would require three main issues to be addressed. These are: (1) data irregularity,(2) permutation-invariance and (3) resampling-invariance. These issues are discussed briefly below.

Irregularity of the point cloud means that the data elements are not organized on a 2D/3D grid, like the data in 2D/3D images. This mainly originates from the pseudorandom nature of recording (sampling) of the external surface of an object, recorded by e.g. a laser scanner. The irregularity results in an ineffective use of convolutional filters for capturing the spatial-local correlation in data (Li et al., 2018), as they work best on organized data.

Permutation-invariance refers to the geometrically unordered presentation of a point cloud. If we present a point cloud by a matrix in which each row contains a point, alternating the order of the rows does not change the data semantics, while it does affect the numerical computation in deep learning architectures.

Resampling-invariance is a property that means random selection of a sufficiently large subset of the points, preserving the global structure of the object captured by the overall point cloud. The IOS data contains tens of thousands of points. The number of points can vary considerably between two scans, or even between different acquisition runs of the same object. Processing such large-scale and variable-size data is challenging for a deep learning model. Hardware limitations (e.g. memory of the GPU) and working with fixed-rank matrices require a resampling stage. However, a naive resampling approach can invoke the loss of important information and is highly application-dependent.

Since 2016, several studies have investigated point cloud analysis by artificial neural networks (ANNs) for object classification/segmentation tasks. *PointNet* (Qi et al., 2017) and *DeepSets* (Ravanbakhsh et al., 2016) are two pioneering works from recent years, based on the multi-layer perceptron (MLP) network, recently followed by other researchers (Le and Duan, 2018; Li et al., 2018). Available deep learning models include some inventive techniques for the joint handling of the first two mentioned issues (i.e. irregularity and permutation invariance), while still addressing the third issue by applying a uniform resampling for fixing the number of points. Although such an approach is sufficient for many applications like object classification (e.g. classifying the chairs vs. tables), it does not preserve the finer details of data which is important for the segmentation tasks (e.g. classifying a point close to the borderline of a tooth and gingiva). This last issue if not addressed, causes significant performance loss in semantic segmentation tasks.

In this paper, we propose an end-to-end learning framework for IOS segmentation based on recent point cloud deep learning models. Our contribution is threefold.

1. To the best of our knowledge, this is the first end-to-end learning study, proposed for IOS point cloud segmentation.

2. We propose a unique non-uniform resampling mechanism, combined with a compatible loss function, for training and deploying a deep network. The non-uniform resampling facilitates the training and deployment of the network on a fixed-size resampled point cloud which contains different levels of spatial resolution, involving both local, fine details and the global shape structure.

3. In addition to a point-wise classification loss, we employ an adversarial loss for empowering the segmentation network to learn the realistic layout of the labeling space and improving the

classification of points by involving the high-level semantics and preserving the valid arrangement of the teeth labels on the dental arch. In contrast to the existing similar approaches, the discriminator network is applied only to the statistics which are computed from the spatial distributions of labels and the predictions. Consequently, only a shallow network is employed as discriminator that facilitates the training.

2. Related work

Related literature has been divided into two parts: conventional IOS segmentation methods and available deep learning solutions for geometric point cloud IOS analysis.

Conventional IOS segmentation approaches: The existing literature on IOS segmentation is extensive and based on conventional computer graphic/vision algorithms. Among the proposed methods, one generic approach is first projecting the 3D IOS mesh on one or multiple 2D plane(s) and then applying standard computer vision algorithms. Afterwards, the processed data is projected back into the 3D space. For example, Kondo *et al.* (Kondo et al., 2004) proposes gradient orientation analysis and Wongwaen *et al.* (Wongwaen and Sinthanayothin, 2010) applies a boundary analysis on a 2D projected panoramic depth images for finding teeth boundaries. Most of other studies are based on curvature analysis (Yuan et al., 2010; Kumar et al., 2011; Yaqi and Zhongke, 2010; Yamany and El-Bialy, 1999; Zhao et al., 2006), *fast marching watersheds* (Li et al., 2007), *morphological operations* (Zhao et al., 2006), 2D (Grzegorzek et al., 2010) and 3D (Kronfeld et al., 2010) active contour (snake) analysis and tooth-target harmonic fields (Zou et al., 2015) for segmenting the teeth and gingiva. Some other works follow a semi-automatic approach by manually setting a threshold(Kumar et al., 2011), picking some representative points (Yamany and El-Bialy, 1999), or interactively involve a human operator for the analysis (Yaqi and Zhongke, 2010; Zhao et al., 2006). Such a method is always limited by having to find the best handcrafted features, the manual tuning of several parameters, and also the inherent limitation of handcrafted CAD systems.

Deep learning approaches: The available deep learning approaches for structured learning on geometric point clouds can be roughly categorized into four types: *feature-based deep neural networks (DNNs)*, *volumetric*, *2D projection* and *point cloud* methods. *Feature-based DNNs* first extract a set of standard shape features (e.g. based on computer graphic algorithms) and then apply a neural network (e.g. a CNN) for feature classification (Guo et al., 2015; Fang et al., 2015). The performance of this approach is limited to the discriminating properties of the handcrafted features (Qi et al., 2017). The *volumetric* approach, first voxelizes the shape and then applies 3D CNN models on the quantized shape into a 3D grid space (Wu et al., 2015; Qi et al., 2016). As expected, the spatial quantization constrains such a method's performance, especially when fine, high-frequency details need to be preserved in shape curvatures for accurate prediction. The *2D projection* approach first renders the 3D data into one/multiple 2D plane(s) and then applies the 2D convolution operator for the 2D-image pixel classification and then the processed data is projected back into the 3D data (Kalogerakis et al., 2017). *Point cloud* deep learning models work directly with raw point clouds (Qi et al., 2017; Ravanbakhsh et al., 2016; Li et al., 2018; Le and Duan, 2018). Each point has some attributes, mainly their 3D coordinates and sometimes other attributes like the normal of a surface they may represent, color, etc. Currently, point cloud deep learning models are a very active research track. This last approach does not suffer from some shortcomings that occur when using handcrafted features, quantization errors or high processing demands, as is the case with earlier mentioned approaches.

In this paper, we have setup our methodology and experiments for teeth semantic segmentation based on the PointCNN model (Li et al., 2018). The PointCNN model is based on a $\chi-Conv$ operator, which weighs and permutes the input points and their corresponding features, prior to processing them by a typical convolution operator. The field of view of each $\chi-Conv$ operator consists of a fixed set of k-nearest neighbour (KNN) points. The outcome of the $\chi-Conv$ operation is the aggregation and projection of KNN point features into a representative set of points, after which a typical convolution is applied to them. The PointCNN has a lower amount of parameters and has been shown to be effective for learning local correlations from point cloud data (Li et al., 2018). This is beneficial, because it is less prone to severe overfitting on a small dataset.

3. Method

The block diagram for our proposed method is shown in Figure 1. In the following section, we will discuss our proposed framework in three parts: preprocessing and data augmentation, non-uniform resampling, and model architecture.

3.1. Data augmentation

The training data are augmented by random 3D rotations, point ordering permutations, adding artificial noise (in the form of jittering) to the positions of each point, and instance dropouts. Here, the dropout of instances means randomly removing all points that belong to a specific tooth from the point cloud in each batch of the training data. This helps the network to learn the labels that may be lacking and do not occur in the training set. The only preprocessing that is applied to the input point cloud is normalization of coordinate information within a scan to have a zero mean and unit variance.

3.2. Non-uniform resampling

Because of the mentioned resampling-invariance property of the point cloud, training a deep learning model on whole set of points of an IOS point would lead to potential issues, such as the variable-rank matrices (the number of points in our IOS datasets may vary in amount between [100k, 310k]) as well as the hardware limitations (such as available memory) for processing of the large-scale point cloud. Applying a patch-classification technique which is common for large-size 2D/3D images, would degrade the quality of results because the extracted patches (i.e. a local subset of points) lack global-structure contents. Furthermore, it would also miss the existing strong dependency between the label of each point and its location in the point cloud. Unfortunately, as we already mentioned, the alternative solution based on uniform resampling does not lead to an accurate analysis of data at its highest available resolution. Recently, various non-uniform resampling methods have been proposed by means of optimization of different metrics that preserve high-frequency contents (Chen et al., 2018; Huang et al., 2013) or local directional density (Skrodzki et al., 2018). However, the effectiveness of using such data abstraction methods on the performance of a deep network cannot be easily established and is in contrast with our interest in designing an end-to-end learning scheme that works directly on the raw data. It is preferable to have such an abstraction of information be performed by the network itself with respect to its objective function. Our proposed non-uniform resampling method is based on the Monte Carlo sampling technique and results in a locally-dense and globally-sparse subset of points for training the deep learning model.

We now state the problem more formally. We assume a matrix representation for the point cloud ($X = [x_1, x_2, \ldots, x_N]$) with N points of which each point has D attributes. The point $x_i \in \mathbb{R}^D$ and the point cloud $X \in \mathbb{R}^{N \times D}$, where $D = 3$ for the 3D geometric points. By introducing a radial basis function (RBF), denoted by $\mathscr{K}$, which is positioned on a randomly chosen point ($x_{fovea} \in X$), the geometrical similarity (spatial distance) to the point x_{fovea} can be measured with a weighted distance metric, as specified in Eq.(1). In accordance with the *foveation* as defined in the work of Ciresan *et al.* (Ciresan et al., 2012), we call this point the *fovea*. The RBF kernel is specified by:

$$\mathscr{K}(x_i, x_{fovea}) = exp\left(-\frac{\left\|x_i - x_{fovea}\right\|^2}{2\sigma^2}\right), \tag{1}$$

where σ is a free parameter that controls the bandwidth (*compactness*) of the kernel. By resampling, we aim to choose a subset Y out of X with M points ($M < N$) that has a dense sampling around the fovea and a sparse sampling for farther locations. According to Monte Carlo sampling, by randomly drawing (with replacement) a point x_i from the set X, we accept to insert such a point into the subset Y, only if $\mathscr{K}(x_i, x_{fovea}) > r_\delta$ is satisfied, otherwise it is rejected. The variable r_δ is a random number from a uniform distribution within the unity interval according to the Monte Carlo technique. This process continues until $M-1$ unique points are accepted. Algorithm 1 in the Appendix shows these steps in detail. Hence, the resampled subset Y has M total points at different levels of granularity (see Figure 2). By random selection of the fovea in every training batch, the model trains on the whole point cloud in its highest available resolution with a fixed number of points. It worths to mention that as the point cloud is normalized to have variance of unity, the uniform-resampling and patch sampling both can be considered as two extreme cases of our proposed algorithm by setting $\sigma \gg 1$ and $\sigma \ll 1$, respectively.

3.3. Model architecture

Our proposed model includes two networks: the *segmentation* network ($\mathscr{S}$) and the *discriminator* network ($\mathscr{D}$). The PointCNN (Li et al., 2018) architecture is used for implementing the $\mathscr{S}$ network. The inputs to the S network are the resampled points and its output is a 17-element vector for each point, which represents the class probability.

Weighted point-wise cross entropy loss: Training the segmentation network by computing an equally weighted loss for each point in the input non-uniform resampled data is not efficient. Since the resampled point set contains various levels of granularity, equally penalizing the output errors for dense and sparse regions prevents the model from optimally adapting its convolutional kernels to capture the fine-detailed content in the data, as the error on sparse points increases relatively equally. Figure 2 shows the uncertainty values for each point, predicted by the network with an equally weighted loss function. As expected, the sparse regions with a lower sampling rate yield a high uncertainty during the learning process because the missing context makes it difficult for the model to perform as accurately as it performs in dense regions. For optimizing the performance of the learning algorithm, we have to trade-off the preservation of the sparse points (which contain the global dental arch structure) and learning of the fine-curvature in point cloud data by parameter tuning. To do so, we apply different weights per point which are computed with the distance metric of the RBF kernel (Eq.(1)). By assuming the posterior probability vector ($\mathbf{P}_i$) for point i, which is computed at the output softmax layer of the segmentation network with the transfer function of $\mathscr{S}$

and its parameters $\boldsymbol{\theta}_{\mathscr{S}}$, the weighted loss value ($\mathscr{L}_p$) for each point i is formulated by:

$$\begin{aligned} \mathbf{P}_i &= [p_{i1}, ..., p_{iL}] = \mathscr{S}(x_i, \boldsymbol{\theta}_{\mathscr{S}}), \quad \text{where} \quad \sum_{j=1}^{L} p_{ij} = 1 \ \text{ with } \ 0 \leq p_{ij} \leq 1, \\ \mathscr{L}_p &= -\sum_{i=1}^{M} w_i \cdot \sum_{j=1}^{L} y_i \cdot \log(p_{ij}), \quad \text{and} \quad w_i = \mathscr{K}(x_i, x_{fovea}). \end{aligned} \tag{2}$$

Here, the y_i represents the one-hot encoded target label for the point i with x_i in 3D coordinates. In our experiments, $L = 17$ and $M = 3 \cdot 10^4$ denote the number of labels and the number of resampled points, respectively.

Adversarial loss Training the segmentation network only by applying a standard pixel-wise (voxel or point-wise) cross-entropy loss function has an important shortcoming. The label of each point in the cloud has a high dependency to the label of its adjacent points. For example, if a point belongs to an incisor tooth, its adjacent points can only belong to the same or another incisor, a canine tooth or to the gingiva, but certainly not belong to a molar tooth. Although such a strong structural constraint exists in the data, it is ignored when the optimization problem is only formulated by Eq. (2). As discussed in (Ghafoorian et al., 2018), the semantic segmentation is inherently not a pixel-based (point-wise) classification problem, hence such a formulation is ill-posed. For improving the higher-level semantic consistencies, Luc *et al.* (Luc et al., 2016) employed an adversarial training in addition to a supervised training of the segmentation network. According to such an approach, a discriminator network provides supervisory signal (feedback) to the segmentation network based on differences between distributions of labels and predictions. Such an effective mechanism was later followed for medical image analysis (Dai et al., 2018; Huo et al., 2018; Kohl et al., 2017; Moeskops et al., 2017; Xue et al., 2018; Yang et al., 2017).

In (Ghafoorian et al., 2018), the authors use a discriminator network to discriminate between the generated labels from a segmentation network and the ground truth labels. Furthermore, they propose using an embedded loss (distance between the features of the hidden layer in the discriminator network) for stability of the training. For point-cloud semantic segmentation, we follow a similar approach, but instead of a heavy training of the discriminator directly on the input space (point cloud and labels) and defining an embedded loss, we first compute two statistical parameters from both the predicted labels and real labels. Afterwards, by training a shallow MLP network as a discriminator, we facilitate the segmentation network's ability to produce a more realistic prediction. The statistics that we used simply consist of the mean and variance of the coordinates of all points with the same label, as given by the segmentation network, which leads to:

$$\hat{\mu}_j = \sum_{i=1}^{M} p_{ij} \cdot x_i \quad \text{and} \quad \hat{\sigma}_j^2 = \sum_{i=1}^{M} p_{ij} \cdot (x_i - \hat{\mu}_j)^2, \quad j = 1, 2, \ldots, L-1 \tag{3}$$

$$\hat{\boldsymbol{u}}_{(L-1)\times 6} = \left[\hat{\mu}_1, \hat{\sigma}_1^2 \parallel \hat{\mu}_2, \hat{\sigma}_2^2 \parallel \ \ldots \ \parallel \hat{\mu}_{L-1}, \hat{\sigma}_{L-1}^2 \right]. \tag{4}$$

Here, the $||$ denotes a vertical vector concatenation (stacking). As mentioned earlier, L denotes the number of labels in the data. The stacked feature set ($\hat{\boldsymbol{u}}$) represents a *soft* computation of the central positions of teeth and their variance (i.e. their soft bounding boxes) in the 3D space, according to the predicted labels (p_{ij}). The statistical mean and variance are computed only for $L-1$ classes of

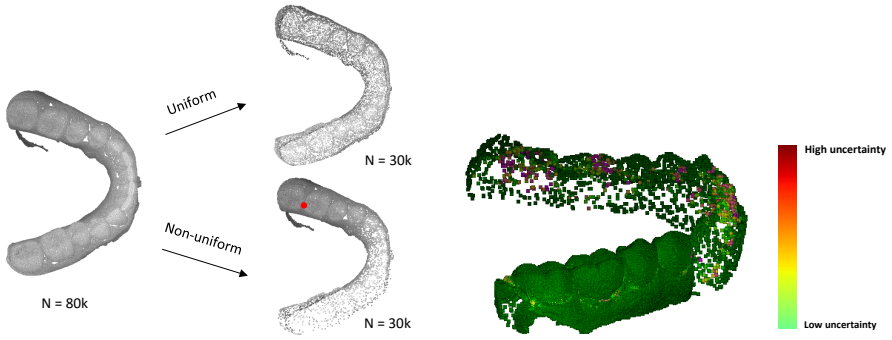

Figure 2: Example of uniform vs. non-uniform resampling (left). The fovea is shown by a red dot. The uncertainty of the prediction for dense and sparse regions (right).

teeth. In computing the high-level semantic features (statistics), we ignore the gingiva class, since its point cloud is almost spread across the whole input space and the applied non-uniform resampling stage alters its resulting statistics severely across training batches. By replacing the p_{ij} values in Eq. 3 with the one-hot encoded values of the ground truth labels (y_i), the counterpart feature-set of $\hat{\boldsymbol{u}}$, denoted by $\boldsymbol{u}$, is obtained. For any absent label in the point cloud, we simply insert a vector consisting of zeros instead. The feature set $\boldsymbol{u}$ represents a *realistic* statistical measurement of the labeled data.

The discriminator network ($\mathscr{D}$) aims to discriminate between feature set $\boldsymbol{u}$ and $\hat{\boldsymbol{u}}$. The network consists of two cascaded parts. The first part estimates an affine transformation and is applied to a input 96-element input vector. The second part consists of 3-layer MLP network which maps the transformed input vector into a scalar value by a sigmoidal activation function at its output node. In effect, the network tries to produce the scalar 1 at its output if the network is applied on $\boldsymbol{u}$, while the scalar 0 should be produced if the network is applied on $\hat{\boldsymbol{u}}$. The architecture of the first part of the network is identical to what is proposed in the PointNet model (Qi et al., 2017), called a *T-Net*. More details about the T-Net can be found in (Qi et al., 2017). In an adversarial setting for training the network *D* and network *S*, the discriminator loss ($\mathscr{L}_{\mathscr{D}}$) for the network $\mathscr{D}$ with parameters $\boldsymbol{\theta}_{\mathscr{D}}$ and an adversarial loss for the network $\mathscr{S}$, can be written as:

$$\begin{aligned}\mathscr{L}_{\mathscr{D}}(\boldsymbol{u},\hat{\boldsymbol{u}};\boldsymbol{\theta}_{\mathscr{D}},\boldsymbol{\theta}_{\mathscr{S}}) &= \mathbb{E}_{\boldsymbol{u}}\big[\log \mathscr{D}(\boldsymbol{u})\big] + \mathbb{E}_{\hat{\boldsymbol{u}}}\big[\log(1-\mathscr{D}(\hat{\boldsymbol{u}}))\big],\\ \mathscr{L}_{Adv}(\hat{\boldsymbol{u}};\boldsymbol{\theta}_{\mathscr{D}},\boldsymbol{\theta}_{\mathscr{S}}) &= \mathbb{E}_{\hat{\boldsymbol{u}}}\big[\log \mathscr{D}(\hat{\boldsymbol{u}})\big].\end{aligned} \tag{5}$$

Hence, the total loss for the segmentation network is a contribution of the losses $\mathscr{L}_p$ in Eq.(2) and $\mathscr{L}_{Adv}$ in Eq. (5). To avoid the need for manual hyper-parameter tuning for the contribution weights (λ) between two loss terms, we follow the work by Kendall *et al.* (Kendall et al., 2017) and involve *adaptive loss weighting*. After initializing $\lambda = [\lambda_1, \lambda_2]$ with a vector of ones, we add the regularization term $\mathscr{R}(\lambda)$ to the total loss function for the segmentation network ($\mathscr{S}$), giving:

$$\mathscr{L}_{\mathscr{S}} = \frac{1}{{\lambda_1}^2}\cdot\mathscr{L}_p + \frac{1}{{\lambda_2}^2}\cdot\mathscr{L}_{Adv} + \mathscr{R}(\lambda), \quad \text{where} \quad \mathscr{R}(\lambda) = {\lambda_1}^2\cdot{\lambda_2}^2. \tag{6}$$

Inference on the whole point cloud: Since the segmentation network is trained on non-uniformly resampled data, for prediction on the whole point cloud we need to extract several subsets of points

according to the non-uniform resampling algorithm. Afterwards, the prediction of all points in the original point cloud is obtained by aggregating all the estimated labels for the extracted subsets. The pseudocode of Algorithm 2 in the Appendix describes this procedure in detail.

4. Experiment and Results

4.1. Data

Our dataset consists of 120 optical scans of dentitions from 60 adults subjects, each containing one upper and one lower jaw scan. The dataset includes scans from healthy dentitions and a variety of abnormalities among subjects. The optical scan data was recorded by a 3Shape d500 optical scanner (3Shape AS, Copenhagen, Denmark). On average, an IOS contains 180k points (varying in range between [100k, 310k]). All optical scans were manually segmented and their respective points were categorized into one of the 32+1 classes by a dental professional and reviewed and adjusted by one dental expert (DAM) with Meshmixer 3.4 (Autodesk Inc, San Rafael CA, USA). Labeling of the tooth categories was performed according to the international tooth numbering standard (FDI). Segmentation of each optical scan took 45 minutes on average, which shows its intensive laborious task for a human.

4.2. Experimental setup

The performance of the model is evaluated by fivefold cross-validation and the results are compared making use of the average Jaccard Index, also known as intersection over union (IoU). On top of the IoU, we report the precision and recall for our multi-class segmentation problem. For computing the precision and recall, each class is treated individually (one-versus-all), as a binary problem and finally the the average scores are reported. Our experiments are partitioned into three parts: (1) benchmarking the performance of the PointCNN in comparison with two other state-of-the-art deep learning models capable of IOS segmentation. These models include PointNet (Qi et al., 2017) and PointGrid (Le and Duan, 2018); (2) Evaluating the impact of applying the non-uniform resampling versus using naive uniform resampling. For the purpose of fair comparison, the number of resampled points are kept identical ($M = 30k$); (3) Evaluating the effectiveness of involving the adversarial loss.

All models are trained utilizing stochastic gradient descent and the Adam learning adaptation technique for 1000 epochs with batch size of one. The initial learning rate is equal to $5e{-}3$, which decreases each $20K$ iterations by a factor of 0.9. We empirically adjust the free parameter of the resampling kernel (see Eq.(1)) to 0.4 (i.e. $\sigma = 0.4$). Since the point cloud is normalized to have a unit variance, we have found that the resampled point cloud by such a chosen setting of σ would encompass at least two teeth in its dense region.

4.3. Results

Table 1 depicts the obtained results from our different experimental setups. Figure 3 in the Appendix shows visualizations of a number of exemplary results from our proposed model. As we can observe from Table1, the PointCNN performs better than two other state-of-the-art models when a naive uniform resampling is applied. This is mostly because of the inclusion of the spatial-correlation information by the $\chi - Conv$ operator in the PointCNN and its lower amount of parameters, which is less prone to overfitting. The PointGrid which samples points inside a predefined grid utilizes

Table 1: Performance of the proposed model within different experimental setups in comparison with state-of-the-art models.

Method			**Metric**			**Exec.time (sec.)**
Network Arch.	Non-uniform	Adv. setting	IoU	Precision	Recall	
PointNet (Qi et al., 2017)	-	-	.76	.73	.65	**0.19**
PointGrid (Le and Duan, 2018)	-	-	.80	.75	.70	0.88
PointCNN (Li et al., 2018)	-	-	.88	.87	.83	0.66
Proposed (I)	✓	-	.91	.90	.87	6.86
Proposed (II)	-	✓	.91	.91	.89	0.66
Proposed (III)	✓	✓	**.94**	**.93**	**.90**	6.86

convolutional operators, but its performance is still limited to the spatial resolution of the spatial quantization grid. The PointNet performance is also constrained, as it omits processing of spatial correlations in the point cloud. With the choice of basing of our method on PointCNN, we show the effectiveness of applying non-uniform resampling and the adversarial loss. The last two techniques improve the results. Finally, incorporating both techniques simultaneously, the highest performance is achieved.

5. Discussion and conclusion

In this paper, we propose an end-to-end learning approach for semantic segmentation of teeth and gingiva from point clouds derived from IOS data. Our segmentation network is based on PointCNN, which has been proposed for point cloud classification/segmentation tasks. For analysis of point clouds in their original spatial resolution (resulting in predictions for all points), we propose a non-uniform resampling mechanism and a compatible loss weighting, based on foveation and Monte Carlo sampling. This resampling approach includes both local, fine-detail information and the sparse global structure of data, which is essential for an accurate prediction of each individual point in absence of a universal coordinate system. Furthermore, by involving the high-level data semantics, through training a discriminator network for learning the realistic layout of labels in data, the results are improved. As a consequence, a heavy post-processing stage (e.g. applying conditional random fields (CRF) on a constructed graph) is not required for incorporating dependencies and locality constraints into the model. By computing the statistics (mean and variance) from spatial distributions of labels and their predictions and feeding them into the discriminator, the adversarial training of the segmentation network is facilitated since for processing such an abstract data only a shallow network can be employed as discriminator. Here, computing the mean and variance of labels and the predictions can be considered generic enough that does not violate the end-to-end learning scheme of the method as using such statistics (operations) is common even within a CNN (e.g. batch normalization operation).

References

Siheng Chen, Dong Tian, Chen Feng, Anthony Vetro, and Jelena Kovačević. Fast resampling of three-dimensional point clouds via graphs. *IEEE Transactions on Signal Processing*, 66(3):666–681, 2018.

Dan Ciresan, Alessandro Giusti, Luca M Gambardella, and Jürgen Schmidhuber. Deep neural networks segment neuronal membranes in electron microscopy images. In *Advances in neural information processing systems*, pages 2843–2851, 2012.

Wei Dai, Nanqing Dong, Zeya Wang, Xiaodan Liang, Hao Zhang, and Eric P Xing. Scan: Structure correcting adversarial network for organ segmentation in chest x-rays. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 263–273. Springer, 2018.

Yi Fang, Jin Xie, Guoxian Dai, Meng Wang, Fan Zhu, Tiantian Xu, and Edward Wong. 3d deep shape descriptor. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2319–2328, 2015.

Mohsen Ghafoorian, Cedric Nugteren, Nóra Baka, Olaf Booij, and Michael Hofmann. El-gan: Embedding loss driven generative adversarial networks for lane detection. *arXiv preprint arXiv:1806.05525*, 2018.

Marcin Grzegorzek, Marina Trierscheid, Dimitri Papoutsis, and Dietrich Paulus. A multi-stage approach for 3d teeth segmentation from dentition surfaces. In *International Conference on Image and Signal Processing*, pages 521–530. Springer, 2010.

Kan Guo, Dongqing Zou, and Xiaowu Chen. 3d mesh labeling via deep convolutional neural networks. *ACM Transactions on Graphics (TOG)*, 35(1):3, 2015.

Hui Huang, Shihao Wu, Minglun Gong, Daniel Cohen-Or, Uri Ascher, and Hao Richard Zhang. Edge-aware point set resampling. *ACM Transactions on Graphics (TOG)*, 32(1):9, 2013.

Yuankai Huo, Zhoubing Xu, Shunxing Bao, Camilo Bermudez, Andrew J Plassard, Jiaqi Liu, Yuang Yao, Albert Assad, Richard G Abramson, and Bennett A Landman. Splenomegaly segmentation using global convolutional kernels and conditional generative adversarial networks. In *Medical Imaging 2018: Image Processing*, volume 10574, page 1057409. International Society for Optics and Photonics, 2018.

Evangelos Kalogerakis, Melinos Averkiou, Subhransu Maji, and Siddhartha Chaudhuri. 3d shape segmentation with projective convolutional networks. In *Proc. CVPR*, page 8, 2017.

Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. *arXiv preprint arXiv:1705.07115*, 3, 2017.

Simon Kohl, David Bonekamp, Heinz-Peter Schlemmer, Kaneschka Yaqubi, Markus Hohenfellner, Boris Hadaschik, Jan-Philipp Radtke, and Klaus Maier-Hein. Adversarial networks for the detection of aggressive prostate cancer. *arXiv preprint arXiv:1702.08014*, 2017.

Toshiaki Kondo, SH Ong, and Kelvin WC Foong. Tooth segmentation of dental study models using range images. *IEEE Transactions on medical imaging*, 23(3):350–362, 2004.

Thomas Kronfeld, David Brunner, and Guido Brunnett. Snake-based segmentation of teeth from virtual dental casts. *Computer-Aided Design and Applications*, 7(2):221–233, 2010.

Yokesh Kumar, Ravi Janardan, Brent Larson, and Joe Moon. Improved segmentation of teeth in dental models. *Computer-Aided Design and Applications*, 8(2):211–224, 2011.

Truc Le and Ye Duan. Pointgrid: A deep network for 3d shape understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9204–9214, 2018.

Yangyan Li, Rui Bu, Mingchao Sun, and Baoquan Chen. Pointcnn. *arXiv preprint arXiv:1801.07791*, 2018.

Zhanli Li, Xiaojuan Ning, and Zengbo Wang. A fast segmentation method for stl teeth model. In *Complex Medical Engineering, 2007. CME 2007. IEEE/ICME International Conference on*, pages 163–166. IEEE, 2007.

Pauline Luc, Camille Couprie, Soumith Chintala, and Jakob Verbeek. Semantic segmentation using adversarial networks. *arXiv preprint arXiv:1611.08408*, 2016.

Pim Moeskops, Mitko Veta, Maxime W Lafarge, Koen AJ Eppenhof, and Josien PW Pluim. Adversarial training and dilated convolutions for brain mri segmentation. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pages 56–64. Springer, 2017.

Charles R Qi, Hao Su, Matthias Nießner, Angela Dai, Mengyuan Yan, and Leonidas J Guibas. Volumetric and multi-view cnns for object classification on 3d data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5648–5656, 2016.

Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 1(2):4, 2017.

Siamak Ravanbakhsh, Jeff Schneider, and Barnabas Poczos. Deep learning with sets and point clouds. *arXiv preprint arXiv:1611.04500*, 2016.

Martin Skrodzki, Johanna Jansen, and Konrad Polthier. Directional density measure to intrinsically estimate and counteract non-uniformity in point clouds. *Computer Aided Geometric Design*, 2018.

Nonlapas Wongwaen and Chanjira Sinthanayothin. Computerized algorithm for 3d teeth segmentation. In *Electronics and Information Engineering (ICEIE), 2010 International Conference On*, volume 1, pages V1–277. IEEE, 2010.

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.

Yuan Xue, Tao Xu, Han Zhang, L Rodney Long, and Xiaolei Huang. Segan: Adversarial network with multi-scale l 1 loss for medical image segmentation. *Neuroinformatics*, 16(3-4):383–392, 2018.

Sameh M Yamany and Ahmed M El-Bialy. Efficient free-form surface representation with application in orthodontics. In *Three-Dimensional Image Capture and Applications II*, volume 3640, pages 115–125. International Society for Optics and Photonics, 1999.

Dong Yang, Daguang Xu, S Kevin Zhou, Bogdan Georgescu, Mingqing Chen, Sasa Grbic, Dimitris Metaxas, and Dorin Comaniciu. Automatic liver segmentation using an adversarial image-to-image network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 507–515. Springer, 2017.

Ma Yaqi and Li Zhongke. Computer aided orthodontics treatment by virtual segmentation and adjustment. In *Image Analysis and Signal Processing (IASP), 2010 International Conference on*, pages 336–339. IEEE, 2010.

Tianran Yuan, Wenhe Liao, Ning Dai, Xiaosheng Cheng, and Qing Yu. Single-tooth modeling for 3d dental model. *Journal of Biomedical Imaging*, 2010:9, 2010.

Mingxi Zhao, Lizhuang Ma, Wuzheng Tan, and Dongdong Nie. Interactive tooth segmentation of dental models. In *Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005. 27th Annual International Conference of the*, pages 654–657. IEEE, 2006.

Bei-ji Zou, Shi-jian Liu, Sheng-hui Liao, Xi Ding, and Ye Liang. Interactive tooth partition of dental mesh base on tooth-target harmonic field. *Computers in biology and medicine*, 56:132–144, 2015.

Appendix

Algorithm 1: Non-uniform resampling

input : point cloud
output: non-uniform resampled point cloud

$X \leftarrow \{x_1, x_2, \ldots, x_N\}$; // whole point cloud
$Y \leftarrow \emptyset$; // initialized empty set
$x_f \leftarrow x \sim X$; // randomly draw one sample as fovea
Function $\mathscr{R}(x_f, X)$ **:**
  while $|Y| < M$; // check the size of Y
  do
    $x_i \leftarrow x \sim X$; // randomly draw sample
    $r_\delta \sim$ uniform $(0,1)$; // draw a random value
    if $\mathscr{K}(x_i, x_f) \geq r_\delta$; // The RBF kernel Eq. 1
    then
      if $x_i \notin Y$ **then**
        $Y \leftarrow x_i \cup Y$; // insert to the subset
      end
    end
  end
  return Y

Algorithm 2: Inference on the whole point cloud

input : point cloud
output: predicted label per point

$X \leftarrow \{x_1, x_2, \ldots, x_N\}$; // whole point cloud
Function $Inference(X)$ **:**
  $U \leftarrow \emptyset$; // initialized an empty set
  $P_X \leftarrow 0_{N \times 17}$; // initialized probability vectors
  while $|U| < |X|$ **do**
    $x_f \sim \{x \in X \mid x \notin U\}$; // select fovea out of the unprocessed points
    $Y \leftarrow \mathscr{R}(x_f, X)$; // non-uniform resampling (Algorithm 1)
    $P_Y = \mathscr{S}(Y, \theta_{\mathscr{S}})$; // prediction of $\mathscr{S}$ Net.
    $\{x_i\} \leftarrow \{x \in Y \mid \mathscr{K}(x_f, x) < \sigma\}$; // only dense region is valid
    $P_X(x_i) = P_X(x_i) + P_Y(x_i)$; // Aggregate the probabilities
    $U \leftarrow \{x_i\} \cup \{U\}$; // Mark as processed
  end
  return $argmax(P_X)$; // labels on whole point cloud

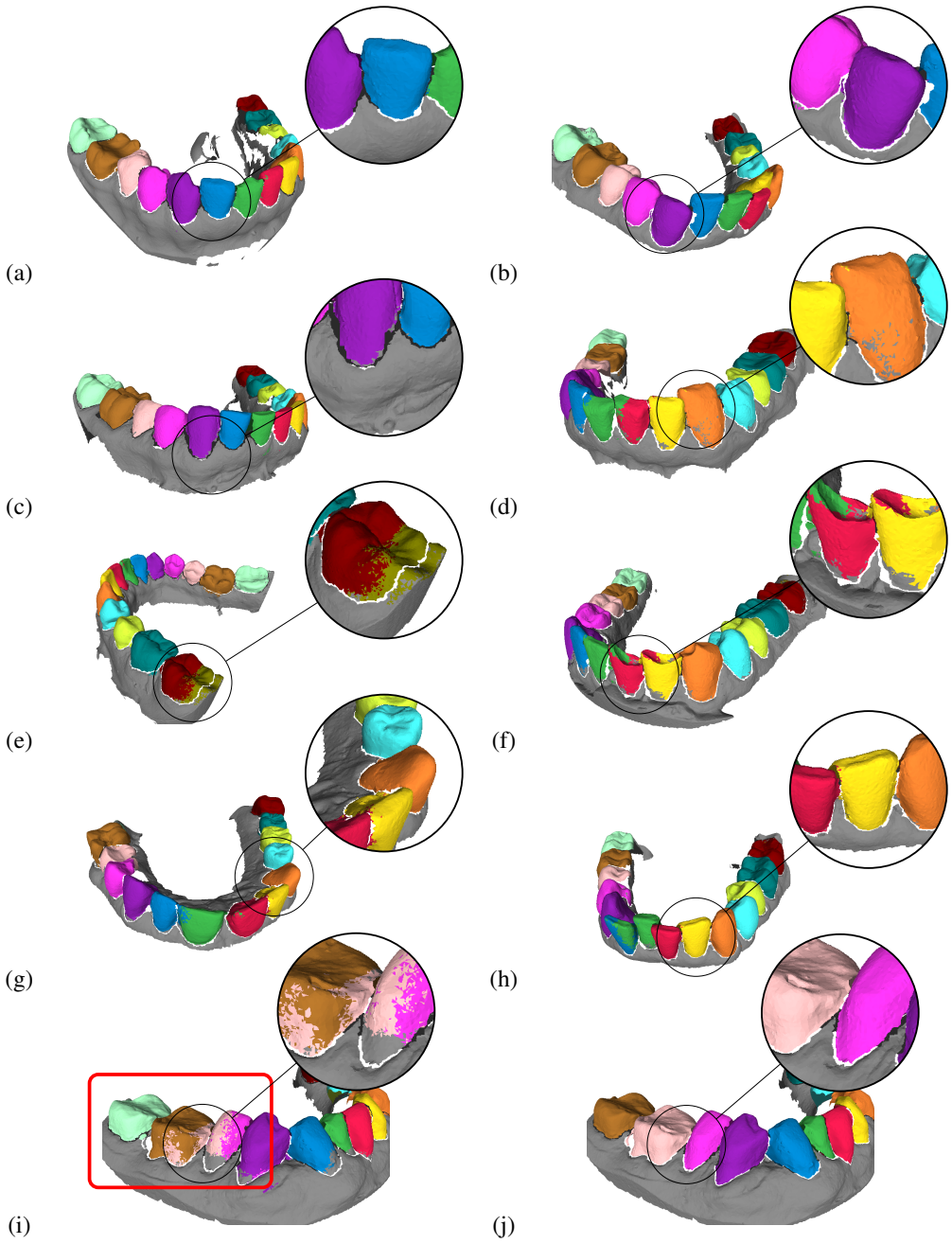

Figure 3: (a-h) Examples of segmentation by our proposed method. (i) Example of a failure case when the adversarial loss is not involved in the training of the segmentation network. The assigned label inside the circle is unrealistic (i.e. invalid). Consequently, the model assigned a set of invalid labels to other neighbouring teeth (inside the red rectangle) by their maximum likelihood. (j) Ground truth.