MeshSNet: Deep Multi-scale Mesh Feature Learning for End-to-End Tooth Labeling on 3D Dental Surfaces

Chunfeng Lian[1], Li Wang[1(✉)], Tai-Hsien Wu[2], Mingxia Liu[1(✉)], Francisca Durán[2], Ching-Chang Ko[3], and Dinggang Shen[1(✉)]

[1] Department of Radiology and BRIC, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
{li_wang,mxliu,dgshen}@med.unc.edu
[2] Department of Oral and Craniofacial Health Sciences, University of North Carolina at Chapel Hill, Chapel Hill, NC 27516, USA
[3] Department of Orthodontics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27516, USA

Abstract. Accurate tooth labeling on 3D dental surfaces is a vital task in computer-aided orthodontic treatment planning. Existing automated or semi-automated methods usually require human interactions, which is time-consuming. Also, they typically use simple geometric properties as the criteria for segmentation, which cannot well handle the high variation of tooth appearance across different patients. Recently, several pioneering deep neural networks (e.g., PointNet) have been proposed in the computer vision and computer graphics communities to efficiently segment 3D shapes in an end-to-end manner. However, these methods do not perform well in our specific task of tooth labeling, especially considering that they cannot explicitly model fine-grained local geometric context of teeth (although only a small portion of dental surfaces but with different shapes and appearances). In this paper, we propose a specific deep neural network (called *MeshSNet*) for end-to-end tooth segmentation on 3D dental surfaces captured by advanced intraoral scanners. Using directly raw mesh data as input, our MeshSNet adopts novel graph-constrained learning modules to hierarchically extract multi-scale contextual features, and then densely integrates local-to-global geometric features to comprehensively characterize mesh cells for the segmentation task. We evaluated our proposed method on an in-house clinic dataset via 3-fold cross-validation. The experimental results demonstrate the superior performance of our MeshSNet method, compared with the state-of-the-art deep learning methods for 3D shape segmentation.

1 Introduction

As a fundamental part of computer-aided-design (CAD) systems for orthodontic treatment planning, accurate tooth segmentation/partition from digitalized dental surface model is a precondition for the analyses and rearrangements of tooth

D. Shen et al. (Eds.): MICCAI 2019, LNCS 11769, pp. 837–845, 2019.
https://doi.org/10.1007/978-3-030-32226-7_93

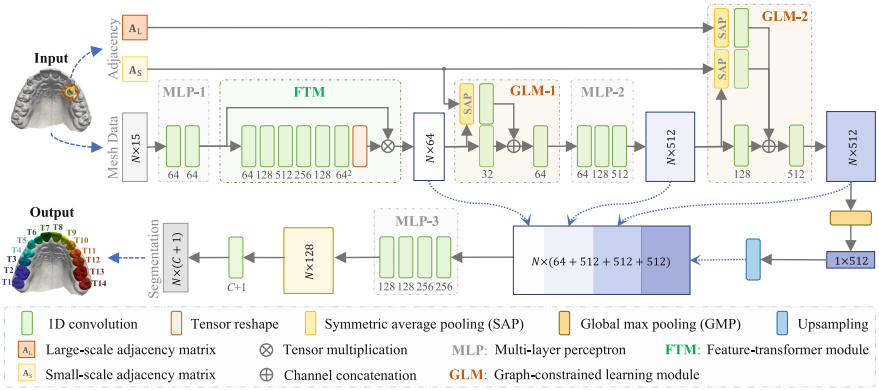

Fig. 1. Illustration of our MeshSNet model, a multi-scale deep neural network to learn high-level geometric features for end-to-end tooth segmentation on 3D dental surfaces.

positions [4]. In clinical orthodontic practice, 3D intraoral scanners (IOS) are becoming widespread for the direct reconstruction of the digital surface model of the dentogingival tissues [5]. Compared with conventional physical impressions, such direct digital impressions are more time-efficient and comfortable for patients, avoiding the potential risk of allergies caused by many constituents of physical impression materials [9]. Automatically segmenting teeth on 3D dental surface is a challenging task, primarily considering that tooth shapes vary dramatically across different subjects and also the patients' teeth usually have abnormal appearances (e.g., neighboring teeth are crowded and misaligned) [12]. The segmentation task becomes even more challenging on the raw dental surface acquired by IOS, since the non-tooth parts (e.g., gingival tissues) usually have significantly irregular shapes and the deep intraoral regions (e.g., the 2nd/3rd molars) may not be perfectly captured by the light source.

Conventional methods for automated or semi-automated tooth segmentation usually project 3D meshes onto 2D images [4] or directly separate 3D meshes according to some preselected geometric properties [14]. Although the ideas are direct and intuitive, most of these conventional methods require time-consuming human interactions, and their performance is sensitive to the variation of tooth appearances [12]. Learning-based shape/geometry analysis has been comprehensively studied in both computer vision and computer graphics communities, which is also potentially applicable to the specific task of tooth segmentation. For example, in [12], hand-crafted geometric features were predefined and reshaped as an image to train a multi-stage convolutional neural network (CNN) for labeling mesh cells on dental surfaces. However, such direct application of CNNs may lead to unstable segmentations, because it ignores the fact that the input geometric features are unordered, i.e., different organizations of them result in different "images". Another potential limitation is that this multi-stage CNN performs different steps independently, which may hamper its practical usage due to system complexity. Recently, a pioneering work of PointNet [7] was pro-

posed for end-to-end 3D shape analysis. Using directly the *raw* geometric data (e.g., the coordinates and normal vectors of point clouds) as input, PointNet learns translation-invariant deep features for shape classification/segmentation, yielding state-of-the-art performance in terms of efficiency and accuracy. The major limitation of the original PointNet is that it ignores the local geometric context, while effectively modeling local structures has been proven to be critical for the success of deep neural networks in fine-grained segmentation tasks. Although some efforts [2,6,11] have been proposed to extend PointNet by including contextual information, they usually coarsely group points into several clusters according to their spatial relationship. Such coarse operations cannot perform well in our specific task of fine-grained tooth segmentation, especially considering that each tooth only takes a very small portion of the entire dental surface.

In this paper, we propose an *end-to-end* deep neural network (called MeshSNet) to learn directly high-level geometric features from the *raw dental mesh* data for automated tooth segmentation, with the schematic diagram shown in Fig. 1. Specifically, our MeshSNet method extends the state-of-the-art PointNet from three aspects: (1) We replace points with mesh cells as input, because mesh cells naturally unite topologically-linked points to show the local structure clearly [1]; (2) We propose *multi-scale graph-constrained learning modules* to explicitly model local geometric context and mimic hierarchical feature learning procedure of CNNs; and (3) We densely fuse cell-wise features, multi-scale contextual features, and translation-invariant holistic features for cell annotation.

2 Method

Input: The input $\mathbf{F}^0$ of our MeshSNet model are the raw mesh surface data with the size of $N \times 15$, where N is the number of mesh cells. Each cell is initially described by a 15-dimensional input. Specifically, apart from the coordinates of the three vertices (9 units) and the normal vector (3 units) of each cell, the relative position (3 units) of each cell with respect to the whole surface is also included to provide supplementary information.

MeshSNet Architecture: As shown in Fig. 1, our MeshSNet follows the architecture of PointNet and employs successive multi-layer perceptrons (MLPs) to extract increasingly higher-level geometric features. Similar to convolutional (Conv) layers in CNNs, the learnable parameters of each MLP in MeshSNet are shared across all input mesh cells. Also, in line with [7], the first MLP (i.e., MLP-1) in our MeshSNet is followed by a feature-transformer module (FTM), which maps all inputs into a canonical feature space to improve the robustness of learned feature representations with respect to potential geometric transformations of input surfaces. Denote $\mathbf{F}^1 \in \mathbb{R}^{N \times 64}$ as the features learned by MLP-1. The FTM predicts an 64×64 transformation matrix $\mathbf{T}$ from $\mathbf{F}^1$, and directly updates the feature matrix as $\hat{\mathbf{F}}^1 = \mathbf{F}^1\mathbf{T}$. Compared with the original PointNet

architecture, the major innovations of our MeshSNet model include: **(1)** graph-constrained hierarchical learning of multi-scale local geometric features, and **(2)** dense fusion of local-to-global features for the segmentation task.

(1) Multi-scale graph-constrained learning: We propose a graph-constrained learning module (GLM) to explicitly capture local geometric context of the input surface. The GLMs (i.e., GLM-1 and GLM-2) are integrated at different stages along the forward path of MeshSNet (i.e., after both FTM and MLP-2), which mimic CNNs to gradually increase the receptive field for learning hierarchical multi-scale contextual features. Specifically, regarding each cell of a 3D mesh as the centroid, we define its neighborhood balls with two different radiuses, and the resulting $N \times N$ adjacency matrices (i.e., $\mathbf{A}_{\mathrm{S}}$ and $\mathbf{A}_{\mathrm{L}}$ for small and large balls, respectively) describe the graph connections between any two cells in the underlying Euclidean space. Based on $\mathbf{A}_{\mathrm{S}}$, GLM-1 in our MeshSNet first applies a graph-based fusion operation (called *symmetric average pooling*, SAP) on $\hat{\mathbf{F}}^1$ (i.e., the output of FTM) to propagate the contextual information (provided by neighboring cells) onto each centroid cell. The resulting feature matrix $\tilde{\mathbf{F}}^1 \in \mathbb{R}^{N\times 64}$ encoding local geometric context has the form of

$$\tilde{\mathbf{F}}^1 = \left(\tilde{\mathbf{D}}_{\mathrm{S}}^{-\frac{1}{2}} \tilde{\mathbf{A}}_{\mathrm{S}} \tilde{\mathbf{D}}_{\mathrm{S}}^{-\frac{1}{2}}\right) \hat{\mathbf{F}}^1, \tag{1}$$

where $\tilde{\mathbf{A}}_{\mathrm{S}} = \mathbf{A}_{\mathrm{S}} + \mathbf{I}$ can be regarded as an adjacency with self-loops, $\tilde{\mathbf{D}}_{\mathrm{S}}^{-\frac{1}{2}} \tilde{\mathbf{A}}_{\mathrm{S}} \tilde{\mathbf{D}}_{\mathrm{S}}^{-\frac{1}{2}}$ is the respective symmetric-normalized adjacency, and $\tilde{\mathbf{D}}_{\mathrm{S}}$ is the diagonal degree matrix. After SAP, both $\tilde{\mathbf{F}}^1$ and $\hat{\mathbf{F}}^1$ are further squeezed by shared-weights 1D Convs with 32 channels. The resulting feature matrices are then concatenated across channels, followed by the fusion by another 1D Conv with 64 channels. Notably, the complete operation of our GLM is in some sense an extension of graph convolutional network [3], e.g., the output $\mathbf{F}^{\mathrm{S1}}$ of GLM-1 has the form of

$$\mathbf{F}^{\mathrm{S1}} = \sigma\Big(\Big\{\sigma\big(\hat{\mathbf{F}}^1\mathbf{W}^1\big) \oplus \sigma\big(\tilde{\mathbf{F}}^1\mathbf{W}^1\big)\Big\}\mathbf{W}^2\Big), \tag{2}$$

where $\sigma\big(\tilde{\mathbf{F}}^1\mathbf{W}^1\big) = \sigma\big(\tilde{\mathbf{D}}_{\mathrm{S}}^{-\frac{1}{2}} \tilde{\mathbf{A}}_{\mathrm{S}} \tilde{\mathbf{D}}_{\mathrm{S}}^{-\frac{1}{2}} \hat{\mathbf{F}}^1\mathbf{W}^1\big)$ is similar to a graph Conv layer [3], $\sigma(\cdot)$ is the ReLU activation, $\oplus$ denotes channel-wise concatenation, and $\mathbf{W}^1$ and $\mathbf{W}^2$ are the learnable weights for 1D Convs with 32 and 64 channels, respectively.

Different from GLM-1, GLM-2 enlarges the receptive field and learns multi-scale contextual features. Specifically, based on Eq. (1), the $N \times 512$ feature matrix from MLP-2 (i.e., $\mathbf{F}^2$) is processed by two parallel SAPs in terms of $\mathbf{A}_{\mathrm{S}}$ and $\mathbf{A}_{\mathrm{L}}$, respectively. The resulting feature matrices and $\mathbf{F}^2$ are then squeezed by shared-weights 1D Convs with 128 channels, which are finally concatenated across channels and fused by another 1D Conv with 512 channels. Notably, although we empirically use only two GLMs in our current implementation, as a general architecture, our MeshSNet can integrate more GLMs along its forward path to learn more scales of contextual features according to task requirements.

(2) Dense fusion of local-to-global features: Following [7], we apply global max pooling (GMP) on the output of GLM-2 to produce the translation-invariant holistic features, aiming to encode the semantic information of the whole dental surface. Different from PointNet that inserts only skip connections between cell/point-wise and holistic features, we assume that the multi-scale contextual features (produced by intermediate GLMs) could provide additional information to comprehensively describe mesh cells. Correspondingly, we densely concatenate the local-to-global features from FTM, GLM-1, GLM-2 and (upsampled) GMP, followed by MLP-3 to yield a $N \times 128$ feature matrix. Based on this matrix, a 1D Conv layer with softmax activation is used to predict a $N \times (C+1)$ probability matrix, with each row denoting the probabilities of the respective cell belonging to specific categories (i.e., C teeth and gingiva).

Implementation and Data Augmentation: As shown in Fig. 1, our MeshSNet model consists of three MLPs (i.e., MLP-1, MLP-2, and MLP-3), one FTM, two GLMs (i.e., GLM-1 and GLM-2), and a final 1D Conv layer to output softmax segmentation probabilities. MLP-1 contains two 1D Convs, both with 64 channels. MLP-2 has three 1D Convs, each with 64, 128, and 512 channels, respectively. MLP-3 contains four 1D Convs, each with 256, 256, 128, and 128 channels, respectively. All 1D Convs in these MLPs and the GLMs are followed by batch normalization (BN) and ReLU activation. To learn the 64×64 feature transformation matrix $\mathbf{T}$ for $\mathbf{F}^1$ (i.e., the output of MLP-1), FTM employs six 1D Convs with 64, 128, 512, 256, 128, and 64^2 channels, respectively, where each of the first five layers is followed by BN and ReLU, while the last layer (without BN and ReLU) is followed by a tensor reshape operation.

Our MeshSNet was implemented using Python based on Keras. It was trained by minimizing the generalized Dice loss [10] using the AMSGrad variant [8] of the Adam optimizer (mini-batch size: 10; number of epochs: 200). To improve the generalization ability of trained networks, we augmented the training and validation sets by combining (1) random rotation, (2) random translation, and (3) random rescaling (e.g., zoom-in/out) of each 3D surface in reasonable ranges. After that, on each training/validation surface (with roughly $10,000$ cells), we randomly sampled 50% cells from each tooth and then randomly sampled the remaining cells from the gingival as the network input (with $6,000$ cells in total). Notably, the combination of all above operations could largely enrich the training set, and also mitigate the imbalanced learning challenge caused by the fact that each tooth only takes a very small part on the whole dental surface. After network training, we directly applied trained networks on unseen test surfaces to predict the corresponding segmentations. That is, in contrast to the training phase, our network can *directly process the whole dental surfaces with varying sizes in the test phase*, which should be a practically meaningful property in practice.

3 Experiments

Dataset and Experimental Setup: The raw dataset studied in this paper consists of 20 maxillary dental surfaces for different subjects acquired by an in-house 3D IOS. The original surfaces roughly contain 100,000 mesh cells, which were downsampled to 10,000 cells while preserving the original topologies. The ground-truth segmentations for $C = 14$ teeth (i.e., from the central incisor to the second molar on both left and right sides) were manually annotated by a dental resident (guided by experienced dentists) on downsampled surfaces.

We performed 3-fold cross-validation on this dataset. In each iteration, one surface was randomly selected from the training set for validation, and the resulting training and validation sets were then enlarged using the data augmentation protocol described in Sect. 2, by simulating 100 "new" surfaces for each training/validation surface. The training/validation inputs (size: $6,000 \times 15$) were then randomly sampled on each surface on-the-fly. Using the same experimental setup, loss function, and optimizer, we compared our **MeshSNet** method with the state-of-the-art **PointNet** approach [7]. For a more comprehensive evaluation, we also designed a dense variant of PointNet (called **PointNet-D**), in which intermediate features were concatenated with mesh-wise and holistic features for the segmentation task. To verify the effectiveness of two essential components (i.e., multi-scale graph-constrained learning and dense fusion of local-to-global features) of MeshSNet, we also compared MeshSNet with its two variants, called **MeshSNet-S** and **MeshSNet-F**, respectively. In MeshSNet-S, the $\mathbf{A}_{\mathrm{L}}$-related SAP and Conv layers were removed from GLM-2, and the respective network can only perform mono-scale local context modeling. In MeshSNet-F, we only fused the mesh-wise and holistic features for MLP-3, by removing the connections from GLM-1 and GLM-2. Based on the ground-truth annotations, the segmentation results were quantitatively evaluated by three metrics, i.e., Dice similarity coefficient (DSC), sensitivity (SEN), and positive prediction value (PPV).

Table 1. Segmentation results (mean ± standard deviation) for all teeth quantified under 3-fold cross-validation, where p indicates the p-value for the statistical significance comparison between our MeshSNet approach and each competing method.

Metric	PointNet	PointNet-D	MeshSNet-S	MeshSNet-F	MeshSNet
DSC	0.781 ± 0.134	0.806 ± 0.121	0.859 ± 0.134	0.894 ± 0.083	$\mathbf{0.938 \pm 0.060}$
	$p = 1.6\text{e}{-10}$	$p = 7.6\text{e}{-8}$	$p = 5.1\text{e}{-4}$	$p = 1.5\text{e}{-3}$	n/a
SEN	0.828 ± 0.167	0.867 ± 0.146	0.882 ± 0.151	0.903 ± 0.100	$\mathbf{0.946 \pm 0.062}$
	$p = 8.1\text{e}{-7}$	$p = 6.3\text{e}{-5}$	$p = 8.5\text{e}{-3}$	$p = 6.5\text{e}{-3}$	n/a
PPV	0.766 ± 0.163	0.772 ± 0.145	0.849 ± 0.141	0.893 ± 0.099	$\mathbf{0.934 \pm 0.077}$
	$p = 6.1\text{e}{-9}$	$p = 3.4\text{e}{-7}$	$p = 5.8\text{e}{-4}$	$p = 2.2\text{e}{-2}$	n/a

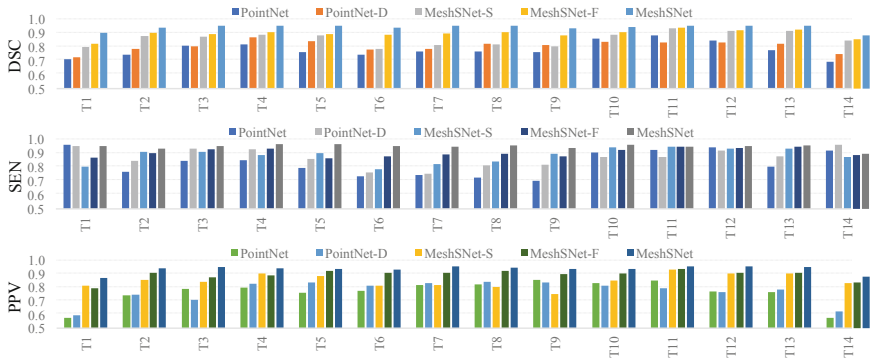

Fig. 2. Segmentation results for each of 14 teeth (i.e., T1–T14) quantified under 3-fold cross-validation, in terms of three evaluation metrics (i.e., DSC, SEN, and PPV).

Results: In terms of the three metrics, the overall segmentation results for all teeth are summarized in Table 1, and the specific segmentation results for each tooth are detailed in Fig. 2. From Table 1, we can have at least three observations. *First*, compared with the state-of-the-art PointNet method, our MeshSNet and its two variants (i.e., MeshSNet-S and MeshSNet-F) led to significantly better results. It suggests that our proposed method could effectively capture and leverage local geometric context to improve the segmentation performance. *Second*, our MeshSNet significantly outperformed its variant MeshSNet-S in terms of all metrics, which implies that explicitly learning multi-scale contextual features is desired for tooth segmentation on dental surfaces, considering that the density of mesh cells may vary across different surfaces and/or different positions. *Third*, our MeshSNet also yielded superior performance than its variant MeshSNet-F. This indicates that, compared with using solely the local and global features, the dense fusion of local-to-global (i.e., cell-wise, multi-scale contextual, and holistic) features could bring additional information for more accurate segmentation. By comparing PointNet-D with PointNet, one could see that the dense fusion strategy also boosts the performance of the original PointNet method.

The per-tooth segmentation results presented in Fig. 2 are consistent with the overall segmentation results summarized in Table 1. From Fig. 2, we can see that our MeshSNet yielded better DSC values than all other competing methods on all teeth (i.e., from T1 to T14), while its variants (i.e., MeshSNet-S and MeshSNet-F) outperformed the state-of-the-art PointNet and its variant (i.e., PointNet-D) on most teeth. These results further verify the effectiveness of our proposed method in the task of automated tooth segmentation on 3D dental surfaces. On the other hand, it is worth noting that, the improvement brought by our MeshSNet method is relatively more significant for the segmentation of *molars* (e.g., T1 and T14), compared with PointNet. For example, our MeshSNet improved the DSC from 0.711 to 0.900 (p-value $< 1e{-}4$), and improved the PPV from 0.575 to 0.867 (p-value $< 1e{-}6$) for segmenting T1. Note that segmenting molars is a very challenging task, because they locate at deep intraoral regions

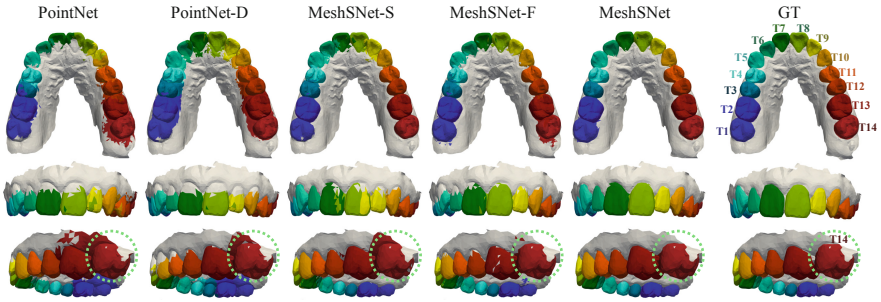

Fig. 3. Segmentations produced by five different methods and the manual ground-truth (GT) annotations for three representative examples.

and might not be completely captured by the light source. These results further suggest the robustness of our proposed method.

In Fig. 3, we visually compare the automated segmentations and ground-truth annotations for three representative examples. From Fig. 3, we can observe that our MeshSNet method has an overall better performance than other competing methods. For example, compared with PointNet and PointNet-D, MeshSNet effectively reduced false positives for segmenting molars (e.g., the first row of Fig. 3), and also reduced false negatives for segmenting central incisors (e.g., the second row of Fig. 3). Especially, our MeshSNet method can more precisely annotate molars that were not completely captured by IOS, which can be observed in green circles in the third row of Fig. 3. Both the visual evaluation in Fig. 3 and the quantitative evaluation in Table 1 and Fig. 2 suggest that our method is potentially useful in practice for automated tooth labeling on dental surfaces.

4 Conclusion

In this paper, we have proposed a deep neural network (called MeshSNet) for end-to-end 3D tooth segmentation on raw dental surfaces acquired by advanced intraoral scanners. Our MeshSNet method integrated novel graph-constrained learning modules to explicitly model the multi-scale local geometric context on mesh surface, and then employed a dense fusion strategy to effectively combine local-to-global features for the comprehensive description of mesh cells. Experimental results on an in-house clinical dataset have demonstrated the superior performance of our proposed method compared with the state-of-the-art deep learning methods for 3D shape segmentation. As the future work, we will integrate trainable post-processing modules (e.g., based on conditional random fields [13]) into our current model to further smooth the segmentations, e.g., by avoiding isolated false positives. In addition, our proposed method should be evaluated on more subjects to further verify its generalization capacity.

References

1. Feng, Y., et al.: MeshNet: mesh neural network for 3D shape representation. In: AAAI (2019)
2. Huang, Q., et al.: Recurrent slice networks for 3D segmentation of point clouds. In: CVPR, pp. 2626–2635 (2018)
3. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. In: ICLR (2017)
4. Kondo, T., et al.: Tooth segmentation of dental study models using range images. IEEE Trans. Med. Imaging **23**(3), 350–362 (2004)
5. Martin, C.B., et al.: Orthodontic scanners: what's available? J. Orthod. **42**(2), 136–143 (2015)
6. Qi, C.R., et al.: PointNet++: deep hierarchical feature learning on point sets in a metric space. In: NeurIPS, pp. 5099–5108 (2017)
7. Qi, C.R., et al.: PointNet: deep learning on point sets for 3D classification and segmentation. In: CVPR, pp. 652–660 (2017)
8. Reddi, S.J., et al.: On the convergence of adam and beyond. In: ICLR (2018)
9. Roberta, T., et al.: Study of the potential cytotoxicity of dental impression materials. Toxicol. Vitro **17**(5–6), 657–662 (2003)
10. Sudre, C.H., et al.: Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In: DLMIA, pp. 240–248 (2017)
11. Wu, W., et al.: PointConv: deep convolutional networks on 3D point clouds. In: CVPR, pp. 9621–9630 (2019)
12. Xu, X., et al.: 3D tooth segmentation and labeling using deep convolutional neural networks. IEEE Trans. Vis. Comput. Graph. **25**, 2336–2348 (2018)
13. Zheng, S., et al.: Conditional random fields as recurrent neural networks. In: CVPR, pp. 1529–1537 (2015)
14. Zou, B., et al.: Interactive tooth partition of dental mesh base on tooth-target harmonic field. Comput. Biol. Med. **56**, 132–144 (2015)