3D Dental Mesh Segmentation Using Semantics-Based Feature Learning with Graph-Transformer

Fan Duan and Li Chen(✉)

School of Software, BNRist, Tsinghua University, Beijing, China
chenlee@tsinghua.edu.cn

Abstract. Accurate segmentation of digital 3D dental mesh plays a crucial role in various specialized applications within oral medicine. While certain deep learning-based methods have been explored for dental mesh segmentation, the current quality of segmentation fails to meet clinical requirements. This limitation can be attributed to the complexity of tooth morphology and the ambiguity of gingival line. Further more, the semantic information of mesh cells which can provide valuable insights into their categories and enhance local geometric attributes is usually disregarded. Therefore, the segmentation of dental mesh presents a significant challenge in digital oral medicine. To better handle the issue, we propose a novel semantics-based feature learning for dental mesh segmentation that can fully leverage the semantic information to grasp the local and non-local dependencies more accurately through a well-designed graph-transformer. Moreover, we perform adaptive feature aggregation of cross-domain features to obtain high-quality cell-wise 3D dental mesh segmentation results. We validate our method using real 3D dental mesh, and the results demonstrate that our method outperforms the state-of-the-art one-stage methods on 3D dental mesh segmentation. Our Codes are available at https://github.com/df-boy/SGTNet.

Keywords: Dental mesh segmentation · Graph-Transformer · Adaptive feature aggregation

1 Introduction

The 3D dental mesh segmentation is aimed to accurately separate the dental mesh into distinct components, namely individual teeth and gums. Therefore stable and accurate 3D dental mesh segmentation plays an essential role in various areas of oral medicine, including orthodontics and denture design, where precise tooth segmentation is of great importance for subsequent procedures and treatments. However, this task is accompanied by notable challenges arising from the inherent limitations in scan accuracy and the presence of considerable noise within the reconstructed 3D dental mesh, consequently leading to the blurring of tooth boundary.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14226, pp. 456–465, 2023.
https://doi.org/10.1007/978-3-031-43990-2_43

To solve these challenges, some methods have been widely explored. Some conventional methods usually utilize specific geometric properties [15,16] like coordinates and normal vectors to perform threshold segmentation. These methods mainly use the pre-defined attributes, making it hard to achieve high-quality results through automated segmentation.

In recent years, many deep learning-based methods have been proposed to perform more accurate automated dental mesh segmentation. Some methods [10,12] pack the point-wise features into 2D image-like inputs which are fed into a multi-layer CNN-like network and some other methods [4,5,9,13] extend the point cloud feature learning frameworks to perform the segmentation. These methods tend to ignore the inherent topology of mesh and thus the quality of the segmentation is not good enough. Some methods [2,8] design a two-stage network consisting of tooth centroid extraction and tooth segmentation. But these methods are not equally effective for models with crowding or missing teeth which are common in real scenes. And they need centroid labels which will incur additional computational cost. Some recent methods like TSGCNet [14] design a two-stream graph convolution-based feature extraction network to extract the features of the C-stream (coordinate) and the N-stream (normal vector) separately and predict the cell/vertex-wise segmentation results according to the concatenated features of two streams. TSGCNet pioneered a new dental mesh feature process paradigm of decoupling the initial feature into C-stream (coordinate) and N-stream (normal vector), and used the graph convolution [11] to consider the topological continuity when updating the features. However, these methods still suffer from the following deficiencies.

First of all, these methods still have some difficulty on the tooth boundary especially for the crowded teeth. Although there always exist individual differences in the shape of teeth from person to person, the teeth at different positions do have distinctive shape priors that distinguish them from other teeth, such as canine teeth and molar teeth. If we can make full use of the shape priors attached to the semantic information of the teeth, these segmentation errors on the boundaries can be decreased greatly. Thus it provides a more promising way to perform the segmentation guided by the semantic information. Secondly, these methods are often confused at the molars since they only use graph convolution to model the dependencies. Therefore a better alternative would be utilizing the local and non-local semantic information at the same time to further enhance the features, that means the long distance dependencies also need to be considered. Lastly, the existing methods always directly perform concatenation on the features from different angles, resulting in the incorrect segmentation on the misaligned teeth. This is due to the fact that the importance of features from different perspectives can vary a lot in different regions. Hence regressing a specific weight and fusing the features with these weight parameters can effectively eliminate the feature imbalance and pay more emphasis on salient features.

To address these issues, we propose a novel semantics-based feature learning network to fully utilize the semantic information and grasp the local and non-local dependencies. We first follow the TSGCNet [14] to decouple the fea-

tures into coordinate domain (C-domain) and normal domain (N-domain) which indicate the spatial and geometric features respectively. And then we design a multi-scale encoder network and at each scale we utilize a coarse classifier which accepts the adaptively fused features from the C-domain and the embedded N-domain to predict a semantic pseudo label. Then with the semantic label we use the graph-transformer module to model the long distance dependencies in the neighbourhood of cells with the minimal semantic distance and perform feature aggregation according to the dependencies. Last but not least, we use the global graph-transformer module on the cross-domain features to further learn the semantic information and fuse them by adaptive feature fusion module before fed into the decoder.

To conclude, our contributions are three-fold: (1) We propose a novel semantics-based feature learning which can fully utilize semantic information to enhance the local and global mesh features. (2) We design a new feature fusion module that obtains global dependencies in C-domain and N-domain to further utilize the semantic information and adaptively fuses cross-domain features. (3) We compare with several recent methods on the real 3D dental mesh collected by the hospital. The OA (Overall Accuracy) and mIoU (mean Intersection over Union) both indicates that we perform superior performance on the 3D dental mesh segmentation task. Extensive evaluations prove that our method significantly outperforms the state-of-the-art one-stage methods.

2 Method

2.1 Overview

Our network mainly consists of a **semantics-based graph-transformer** module and an **adaptive cross-domain feature fusion** module, as shown in Fig. 1. We follow the TSGCNet [14] to fetch the initial cross-domain features as two $N \times 12$ matrices, but our N-domain features serve as the embedding domain instead. The semantics-based graph-transformer is a multi-scale encoder and at each scale we aggregate the features in the neighbourhood of cells with minimal semantic distance provided by the coarse semantics prediction module. The N-domain features are embedded through the adaptive feature fusion module to extract more accurate semantic information. And for the concatenated features from the different scales, we perform the global graph-transformer block respectively in the two domains and fuse them through the same adaptive fusion strategy for the subsequent cell-wise segmentation.

2.2 Semantics-Based Graph-Transformer

The semantics-based graph-transformer module aims to generate the multi-scale cell-wise feature vectors in C-domain embedded with N-domain which can represent the geometric features of dental mesh at different positions accurately and discriminatively. We denote the initial feature vectors extracted from the dental

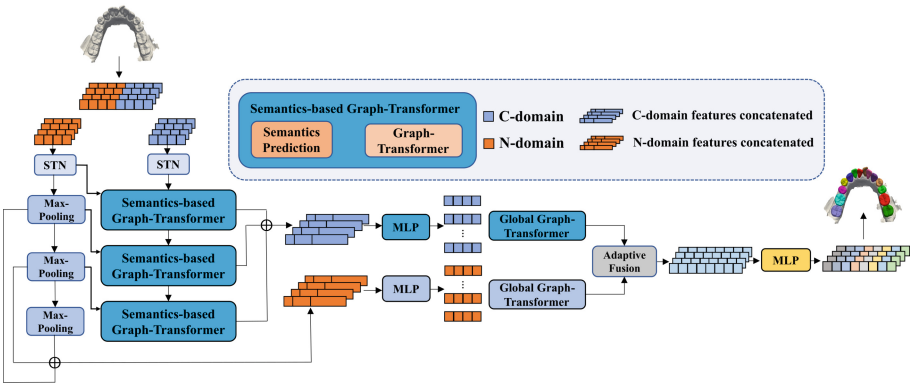

Fig. 1. An overview of the proposed framework.

mesh as a $N \times 24$ matrix, N is the number of the cells and the 24-dimensional vector consists of 12-dimensional relative coordinates and 12-dimensional normal vectors. Then through a normal STN module [3], we make the C-domain and N-domain space invariant due to the fact that the position and orientation of dental mesh can be various. Formed as two domains of $N \times 12$ matrices which represent the spatial position and the geometric features respectively, we perform semantic prediction on the C-domain with the N-domain features embedded to generate a pseudo semantic label $L = \{l_1, l_2, ..., l_n\}$ for each cell. And then according to the semantic information, we use the graph-transformer for each cell in its local neighbourhood where the cells have the minimal semantic differences and update their features to make the difference between the cells with different labels greater. For N-domain, in a geometric sense, they can help classify the semantic label and enhance the local features, hence we mainly upsample it to adapt to the different scales of C-domain without any other modifications.

Semantics Prediction. The semantic prediction is mainly used to generate a pseudo cell-wise label for each cell which can effectively extract the semantic information. Denote the C-domain feature as C that has a shape of $N \times k$, and the N-domain feature as N that has a shape of $N \times k$, and we regress a C-domain weight and a N-domain weight which indicates the weights of the domain fusion. So we have the cross-domain features F adaptively fused as:

$$F_j = c_j \cdot C_j \oplus n_j \cdot N_j \tag{1}$$

where $\oplus$ is the channel-wise concatenation, and j indicates the layer of the semantics-based graph-transformer modules, and c_j, n_j is the adaptive weights of C-domain and N-domain in layer j, and C_j, N_j is the output from the previous layer. And after that we perform a simple MLP to generate the final pseudo semantic cell-wise label formed as:

$$L_j = \mathbf{max}(\mathbf{softmax}(\mathbf{MLP}(F_j))) \tag{2}$$

where **softmax** can get the probability that the cells belong to each class and **max** outputs the index of the maximum value. Thus we have the cell-wise pseudo semantic label which can be used to make the difference between the features of cells belonging to different categories greater.

Graph-Transformer. The graph-transformer is composed of a semantic KNN and a Transformer Encoder Block. For the input C-domain, we first construct a KNN graph based on the semantic biased Euclidean distance formed as:

$$Distance(cell_i, cell_j) = Euclidean(cell_i, cell_j) + Semantic\_Dist(cell_i, cell_j) \tag{3}$$

where $cell_i$ and $cell_j$ indicate the ith cell and jth cell and the $Semantic\_Dist$ function measures the difference of the two cells which is formulated as:

$$Semantic\_Dist(cell_i, cell_j) = \begin{cases} 0, \, l_i = l_j \\ \lambda, \, l_i \neq l_j \end{cases} \tag{4}$$

where λ is a positive parameter that can be set according to the specific task and l_i indicates the pseudo label of the ith cell. Then we perform a transformer encoder block on each cell to get the local dependencies which can enhance the local features belong to each class. The attention we used is a standard multi-head attention, and we set the query and value as the matrices of neighbour features of the cells and the key as the distance matrix between cells and their neighbours.

2.3 Adaptive Cross-Domain Feature Fusion

The adaptive cross-domain feature fusion module aims to fuse the C-domain and N-domain features for the cell-wise segmentation. Through the above semantics-based graph-transformer module we have obtained the accurate multi-scale C-domain features embedded by the N-domain features. Then we need to fuse the features to integrate the spatial information and the geometric information of the cells. Therefore, we first perform the concatenation on the features of different scales and use MLP to fuse the multi-scale features together which can balance the features at different scales. This process is formulated as:

$$\begin{cases} F_c = \mathbf{MLP}(F_c^1 \oplus F_c^2 \oplus F_c^3) \\ F_n = \mathbf{MLP}(F_n^1 \oplus F_n^2 \oplus F_n^3) \end{cases} \tag{5}$$

where $\oplus$ is the channel-wise concatenation while c and n represent C-domain and N-domain features. Then through a global graph-transformer block, we just use the standard multi-head attention on the C-domain and N-domain respectively, so as to capture long distance dependencies and have global knowledge of the semantic information.

Since the learnable weights can fuse the cross-domain features adaptively, we use the same cross-domain feature fusion strategy similar to that we used in the semantics prediction module. Further more, we use a single MLP to generate a

feature mask and perform the dot product on the feature and the mask which is formulated as:

$$\hat{F} = \mathbf{MLP}(F) \odot F \tag{6}$$

where $\odot$ is the element-wise multiplication and F is the fused cross-domain features for segmentation.

3 Experiments and Results

3.1 Implementation Details

Our network is implemented with PyTorch 1.11.0 on four NVIDIA GeForce RTX 3090 GPUs. The input meshes are all downsampled to 12000 cells. And in the training process, we optimize the network through minimizing the cross-entropy loss which is very commonly used in the segmentation task. The learning rate was empirically set as 10^{-3}, and reduced by 0.2 decay every 20 epochs.

3.2 Dataset

Our dataset consists of 200 3D dental meshes obtained by an intraoral scanner on real orthodontic patients from hospital and each raw mesh contains even more than 150,000 cells, so we down sample the raw mesh to 12000 cells while preserving the surface topology. We randomly split the whole dataset as a training set with 160 meshes and a testing set with 40 meshes. And we segment the raw mesh into 14 teeth and gums, following the FDI World Dental Federation notation [1]. This means that each input mesh has 15 labels. For convenience, we do not distinguish between the maxillary teeth and mandibular teeth, and treat the teeth on the opposite side (from maxillary and mandibular teeth respectively) as the same class. And we augment the training set by the random translation between $[-10, 10]$, the random rotation between $[-\pi, \pi]$ and the random scaling between $[0.8, 1.2]$.

For evaluation, overall accuracy (OA) and mean intersection over union (mIoU) are adopted for quantitative comparison.

3.3 Comparing with SOTA Methods

We compare our network against five recent methods, including PointNet++ [7], DGCNN [11], PVCNN [6], MeshSegNet [5] and TSGCNet [14]. For a fair comparison, we utilize the public implementations of compared methods to fine-tune their network for generating their best segmentation results. All the methods are trained for 200 epochs.

The segmentation results are shown in Table 1. All the metrics show that our method outperforms the other methods a lot. Specifically, the TSGCNet [14] is the state-of-the-art one-stage method of the 3D dental mesh segmentation task which pioneered a two-stream graph convolution network to extract the features more accurately and TSGCNet [14] improve the segmentation performance on

Table 1. The segmentation results comparing with five methods on OA and mIoU. Gum-mIoU and teeth-mIoU indicates the mIoU computed on the gums and teeth respectively.

	Method	OA	mIoU	gum-mIoU	teeth-mIoU
Point-based	PointNet++ [7]	87.18	0.708	0.867	0.697
	DGCNN [11]	91.56	0.793	0.927	0.784
	MeshSegNet [5]	93.15	0.825	0.931	0.816
Voxel-based	PVCNN [6]	94.54	0.849	0.934	0.843
Graph-based	TSGCNet [14]	95.63	0.890	0.942	0.886
	Ours	**96.97**	**0.921**	**0.956**	**0.918**

the teeth greatly compared to MeshSegNet [5]. Compared with TSGCNet [14], we still increase the OA and m-IoU on the 3D dental mesh segmentation task by 1.34% and 3.1% respectively.

We also perform the qualitative experiments to further evaluate the segmentation results in Fig. 2. From the visualization results, we can find out that our method also outperforms other methods. In particular, we present some complicated cases where there may exist wisdom teeth, the crowded arrangement or the misplaced teeth. In the first row, the raw mesh has a total of 16 teeth which is different from our setting, and in this case, other methods tend to merge some of the small teeth which will cause confusion, while with semantic information, our method successfully labels all the teeth except the two wisdom teeth. In the second and third row where there exist misplaced teeth and worn teeth, we can see that the previous methods cannot perform segmentation correctly. But with adaptive feature fusion which balances the coordinate features and normal vector features, our method can segment the worn teeth accurately as well as the misplaced teeth guided by the semantic information. And the fourth row demonstrates that in terms of extracting the local geometric features, we can also achieve the superior performance. This qualitative experiment further suggests that our design is more effective and accurate on this segmentation task.

3.4 Ablation Study

We evaluate the effectiveness of semantics prediction, graph-transformer and adaptive feature fusion as three critical components of our method. We perform the evaluation by excluding one of these critical components each time. Specifically, when we remove the semantics prediction module, we will use a naive KNN to construct the KNN graph. When we remove the graph-transformer module, we replace it with max-pooling layers. In the absence of the adaptive feature fusion module, we directly concatenate the features from C-domain and N-domain for cross-domain feature fusion. The results of the ablation study are presented in Table 2. It turns out that semantics prediction, graph-transformer and adaptive feature fusion all bring performance improvement on the 3D den-

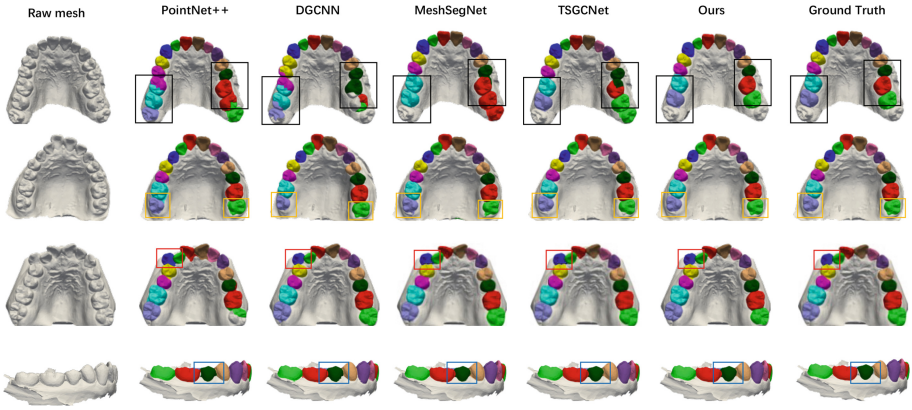

Fig. 2. Visualization of segmentation results comparing with four methods along with the raw dental meshes and ground-truth labels.

tal mesh segmentation task. And we can see that although semantics prediction module improves a little in metrics, it is primarily attributed to the fact that the segmentation errors focus on the boundary cells and the number of such errors is relatively small. But the impact of these boundary cells on the overall segmentation quality is significant.

Table 2. Ablation study of critical components of our method.

Method	OA	mIoU
Ours	**96.97**	**0.921**
Ours w/o semantics prediction	96.52	0.916
Ours w/o graph-transformer	96.18	0.911
Ours w/o adaptive feature fusion	96.05	0.909

4 Conclusion

We propose a novel semantics-based feature learning to make full use of local and unlocal semantic information to enhance the features extracted. This architecture can decouple the spatial and geometric features into C-domain and N-domain and embed the N-domain into C-domain to further utilize the semantic pseudo label to perform the local graph-transformer module. Lastly we fuse the features from spatial and geometric domain adaptively by using the global graph-transformer module and adaptive feature fusion module. The effectiveness of our proposed method is evaluated on the real dental mesh from real orthodontic patients. In the future work, we will try to utilize the feature decoupling and fusing strategy in other segmentation tasks.

Acknowledgement. This research was supported by the National Natural Science Foundation of China (Grant No. 61972221).

References

1. EN ISO 3950:2009: Dentistry-designation system for teeth and areas of the oral cavity (2009)
2. Cui, Z., et al.: TSegNet: an efficient and accurate tooth segmentation network on 3D dental model. Med. Image Anal. **69**, 101949 (2021)
3. Jaderberg, M., et al.: Spatial transformer networks. In: Advances in Neural Information Processing Systems 28 (2015)
4. Lian, C., et al.: MeshSNet: deep multi-scale mesh feature learning for end-to-end tooth labeling on 3D dental surfaces. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11769, pp. 837–845. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32226-7_93
5. Lian, C., et al.: Deep multi-scale mesh feature learning for automated labeling of raw dental surfaces from 3D intraoral scanners. IEEE Trans. Med. Imaging **39**(7), 2440–2450 (2020)
6. Liu, Z., Tang, H., Lin, Y., Han, S.: Point-voxel CNN for efficient 3D deep learning.In: Advances in Neural Information Processing Systems 32 (2019)
7. 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 30 (2017)
8. Qiu, L., Ye, C., Chen, P., Liu, Y., Han, X., Cui, S.: DArch: dental arch prior-assisted 3D tooth instance segmentation with weak annotations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 20752–20761 (2022)
9. Sun, D., et al.: Automatic tooth segmentation and dense correspondence of 3D dental model. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12264, pp. 703–712. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59719-1_68
10. Tian, S., Dai, N., Zhang, B., Yuan, F., Yu, Q., Cheng, X.: Automatic classification and segmentation of teeth on 3D dental model using hierarchical deep learning networks. IEEE Access **7**, 84817–84828 (2019)
11. Wang, Y., Sun, Y., Liu, Z., Sarma, S.E., Bronstein, M.M., Solomon, J.M.: Dynamic graph CNN for learning on point clouds. ACM Trans. Graph. (TOG) **38**(5), 1–12 (2019)
12. Xu, X., Liu, C., Zheng, Y.: 3D tooth segmentation and labeling using deep convolutional neural networks. IEEE Trans. Vis. Comput. Graph. **25**(7), 2336–2348 (2018)
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, pp. 557–571. PMLR (2019)

14. Zhang, L., et al.: TSGCNet: discriminative geometric feature learning with two-stream graph convolutional network for 3D dental model segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6699–6708 (2021)
15. 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, pp. 654–657. IEEE (2006)
16. 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. Comput. Biol. Med. **56**, 132–144 (2015)