IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 15 April 2024, accepted 16 May 2024, date of publication 20 May 2024, date of current version 5 June 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3402950

RESEARCH ARTICLE

Mesh Segmentation for Individual Teeth Based on Two-Stream GCN With Self-Attention

SHI-JIAN LIU[1,2], CHAO-MING KANG[1], FENG-HUA HUANG[2], (Member, IEEE), AND ZHENG ZOU[3]

[1]Fujian Provincial Key Laboratory of Big Data Mining and Applications, Fujian University of Technology, Fuzhou 350118, China
[2]Fujian Key Laboratory of Spatial Information Perception and Intelligent Processing, Yango University, Fuzhou 350015, China
[3]College of Computer and Cyber Security, Fujian Normal University, Fuzhou 350117, China

Corresponding author: Feng-Hua Huang (fhhuang@ygu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 62172095, in part by Fujian Provincial Department of Science and Technology under Grant 2022J01932, in part by Fujian Provincial Department of Education under Grant JAT210283 and Grant JAT220052, in part by the Open Project of Fujian Key Laboratory of Spatial Information Perception and Intelligent Processing Yango University under Grant FKLSIPIP1003, and in part by Hunan Provincial Department of Science and Technology under Grant 2024JJ7549.

ABSTRACT Dental triangular mesh is widely used in computer-aided oral medicine. Different from regular data such as digital images, the structure of triangular mesh is more complex, and traditional operations such as convolution cannot be directly applied. Therefore, the segmentation of patients' personalized teeth from mesh through deep learning is a hot topic in the current research field. Recently, a method named TSGCN presented the idea of learning coordinate features and normal features through a two-stream architecture based on Graph Convolutional Networks (GCN), which further improved the performance compared with other methods. However, its ability to extract and process global features still can be strengthened. To this end, a method named TSGCN-SA is proposed, whose core idea is to introduce the self-attention (SA) mechanism into TSGCN. Specifically, two SA modules are introduced, the first one is used to improve the global feature extraction ability in the coordinate stream. The second one plays an important role in the adaptive contribution adjustment of each stream during the feature fusion. Experiments based on the public dataset named 3DTeethSeg show that TSGCN-SA is superior to SOTAs in terms of segmentation performance due to the proposed SA modules, and the proposed method is competent in the task of individual tooth mesh segmentation.

INDEX TERMS Tooth, triangular mesh, deep learning, self-attention, segmentation.

I. INTRODUCTION

Triangular mesh is one of the most common ways to represent three-dimensional (3D) shapes in the digital world, which is widely used in computer-aided dental medicine. For example, a mesh representing a jaw is depicted in Fig. 1, which can be semantically divided into the teeth area (marked in colors) and the non-teeth area. Obtaining personalized teeth from the dental mesh through segmentation is of great significance for clinical medical applications such as computer-aided orthodontics [1], dental implants [2], and surgical planning [3].

The associate editor coordinating the review of this manuscript and approving it for publication was Junhua Li.

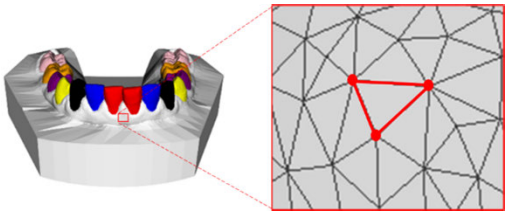


FIGURE 1. Demonstration of the dental mesh.

Although the previous works of this paper [4], [5] have demonstrated superior performance compared to methods of the same period, the way of solving the Laplace harmonic field with manually specified constraints has become its main bottleneck, especially in the current season that most of the

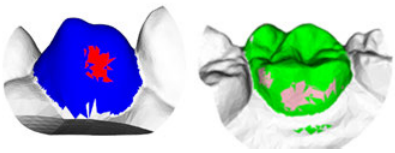


FIGURE 2. Typical results of tooth segmentation using TSGCN.

successful methods rely on deep learning to extract pattern features from big data.

Unlike regular data such as digital images, the structure of mesh is far more complex, and traditional convolution cannot be directly applied. Therefore, dental mesh segmentation via deep learning is one of the cutting-edge research.

Deep learning methods that deal with mesh can be divided into three categories in terms of basic primitives. As shown in the inset of Fig. 1, a mesh consists of many triangular faces, which can be determined by three vertices and the edges between them as well. In other words, vertices, edges, and faces are the basic primitives of a mesh.

MeshCNN [6] is a representative method based on edge primitives. Based on the rule that any edge of a 2-manifold mesh has 4 adjacent edges, MeshCNN successfully extends the convolutional neural network (CNN) from the image domain to the mesh domain by defining a new convolution operation with fixed-size convolution kernels. The definition of pooling is also introduced in their method using mesh simplification techniques.

PointNet++ [7] and MeshSegNet [8] are representatives of the vertex-based (or point-based) method and face-based method respectively. One of the differences between them and MeshCNN is that MeshCNN uses the mesh structure to define the adjacency relationships of basic primitives, while they use the KNN algorithm [9], which helps the aggregation of local features faster.

Compared to the vertex-based method, a major advantage of the face-based method lies in having more training raw features per sample. For example, in terms of spatial coordinate features, the number of feature channels used by the vertex-based method and the face-based method are 3 (i.e., the x, y, z values of 1 vertex) and 12 (i.e., the x, y, z values of 4 points, including the 3 vertices and 1 central point of a face) respectively. It is well-known that richer input features usually indicate greater potential for training a high-performance model. Therefore, deep learning methods based on face primitives are the current research trend.

In terms of methodology, deep learning models used for mesh processing range from CNN [6], [7], [10] to Graph Convolutional Networks (GCN) [8], [11], [12]. Recently, a method named TSGCN [12] presented the idea of learning coordinate and normal features through a two-stream architecture based on GCN. Although TSGCN outperforms the others, experiments have shown that there are often situations as shown in Fig. 2 where small misclassified regions are surrounded by correctly classified regions (e.g., the red part among the blue part on the left of Fig. 2) according to the connectivity marked by colors. This is a sign of models that lack global contextual understanding.

Speaking of global contextual understanding, the Transformer and its self-attention mechanism are well acknowledged. Since the Transformers have become the dominant method in Natural Language Processing (NLP), researchers are devoted to extending it to the field of computer vision. Swin Transformer [13] is such an example, while Point Transformer [14] is devoted to applying Transformers for point clouds. These studies demonstrate that self-attention strategies are capable of effectively extracting global features and benefits for system performance.

Inspired by TSGCN and the Transformers, this paper introduces the self-attention (SA) mechanism into TSGCN for individual tooth mesh segmentation. Therefore the proposed method is named TSGCN-SA by us. The main contributions are as follows:

1) A self-attention architecture named SA-I is proposed. The convolutional layer in the original TSGCN is replaced by SA-I to improve the global contextual understanding ability of the system.
2) To achieve an adaptive balancing of two-stream outputs according to their contributions, a self-attention layer named SA-II is introduced for better cooperation with each branch stream.
3) An automatic dataset construction method and an intelligent human-machine interface are proposed, which ensure the region of interest (ROI) determination easily, intuitively, and robustly.

II. RELATED WORK

A. INDIVIDUAL TOOTH SEGMENTATION

Given a dental mesh, individual tooth segmentation (ITS) refers to segmenting one target tooth each time on the fly [4]. In contrast, whole tooth segmentation (WTS) takes a dental mesh and outputs every tooth on it once and for all.

A major challenge for WTS is that computational resources are highly required. As we know, the scale of a triangle mesh can be measured by the number of faces. Taking the public dental dataset 3DTeethSeg [15] for example, the data scale in 3DTeethSeg ranges from 100,000 faces to 350,000 faces. When feeding TSGCN with a mesh of 130,000 faces, the GPU memory requirement reaches 65 GB, which is far higher than the amount that a mainstream graphics card can offer.

Mesh simplification is a common way to deal with the above problem [10], [11], [12] which will lead to a second challenge: topology or geometry abnormality [16]. There exist algorithms that can guarantee topology-preserving simplification, but the shape will be modified more or less. For example, Fig. 3 depicts the simplification result (18,000 faces) of the mesh (around 180,000 faces) shown in Fig. 1. As the colored annotation shows, the segmentation ground truth (GT) becomes quite chaotic near the ideal boundaries. In other words, the boundary features, which are critical for deep learning based segmentation, may be lost during the simplification.

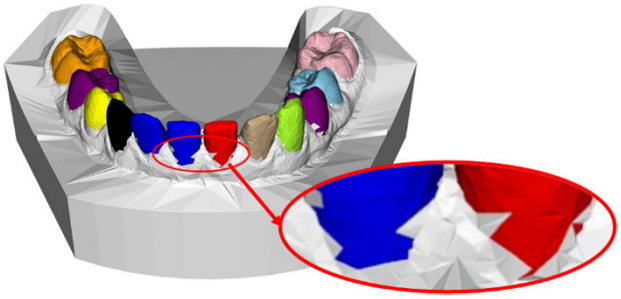


FIGURE 3. The simplification result of data is shown in Fig. 1.

The above challenges could be relieved in ITS because only a local part (i.e., ROI) that contains the target tooth is required compared to the whole dental mesh. Based on this idea, TSegNet [17] first trains a deep learning network to predict the center points of each tooth, and then uses a cascaded network to perform the ITS within the ROI defined by $N/4$ points around each center point, where N is the number of mesh vertices. This paper shares the same local strategy, but we use isotropic spherical cuts to obtain the ROIs, and an intelligent human-machine interface is involved during the inference phase.

B. MULTI-BRANCH FEATURE LEARNING

The essence of deep learning is feature learning. Raw features for meshes are coordinate and normal vectors in general. When building a network, these features can be concatenated and trained undifferentiated [11].

Another strategy is to train different types of features in separate network branches [12]. Experimental results show that the system performance indeed benefits from the personalized multi-branch feature learning scheme. Inspired by this idea, the proposed method also uses a two-stream network architecture. Specifically, we optimized both the branch structure and the fusion behavior in TSGCN.

C. SELF-ATTENTION

Transformer and self-attention have a great reputation in NLP and become acknowledged in the field of 3D shape analysis recently. Generally speaking, self-attention operations can be divided into two categories.

Let x_i represents the *ith* element in the input feature X, the first class can be described by Eq. (1).

$$y_i = \sum_{x_j \in N(i)} \left| \alpha\left(x_i\right)^T \beta\left(x_j\right) \right| \gamma\left(x_j\right) \tag{1}$$

where $N(i)$ indicates the set of features to be aggregated to form a new feature y_i, α, β and γ represents three learnable feature transformations, and $|*|$ is a normalization operation. Since the attention weight in Eq. (1) is a scalar, this one is called scalar attention [18]. The details of scalar attention are shown graphically in Fig. 4(a).

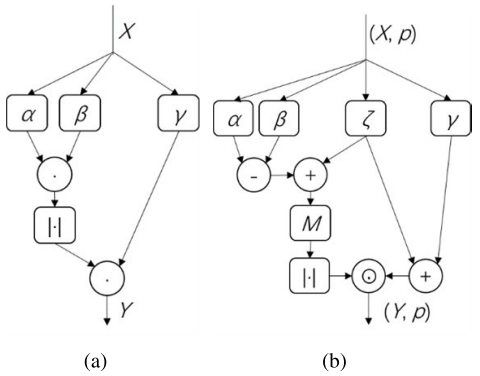


FIGURE 4. The Transformer layer proposed in point transformer. (a) Scalar attention, (b) Vector attention.

The second class is known as vector attention [19], as shown in Eq. (2), because the attention weight is a vector.

$$y_i = \sum_{x_j \in N(i)} \left| M\left\{R\left[\alpha\left(x_i\right), \beta\left(x_j\right)\right] + \zeta\right\} \right| \odot \gamma\left(x_j\right) \tag{2}$$

where R, ζ and M are functions that represent a relationship, positional encoding, and mapping respectively, $\odot$ is the Hadamard product.

Point Transformer [14] adopts the vector attention scheme, and uses linear projection to implement the α, β and γ. In addition, R, M and ζ are defined as $\alpha(x_i) - \beta(x_j)$, multi-layer perceptron (MLP) and $\theta(p_i - p_j)$ respectively, where θ is an MLP with 2 linear layers and 1 ReLU layer, p_i and p_j are the coordinates of the *ith* and *jth* points. It is worth mentioning that Point Transformer uses $\gamma(x_j) + \zeta$ to replace $\gamma(x_j)$ that shown in Eq. (2) because the authors found this modification can improve the accuracy. The graphical details of the SA proposed by Point Transformer are shown in Fig. 4(b).

Other than Point Transformer which builds on pure self-attention modules and is proposed to deal with point clouds, we let the SA work with GCN to enable the cooperation of global and local features for mesh segmentation based on faces.

III. PROPOSED METHOD

Since the core idea of the work is to introduce the self-attention (SA) mechanism into TSGCN, the proposed method is named TSGCN-SA by us, whose architecture is illustrated in Fig. 5.

A. PROCEDURES

As Fig. 5 shows, TSGCN-SA takes the ROI of the target tooth as input and outputs the segmentation result. Specifically, the normal and the coordinate vectors from the ROI are trained in two streams separately. The coordinate stream (C-stream) is stacked with three blocks, which consist of an SA-I layer and a graph attention layer. The output of each block is merged by channel concatenation. The normal stream (N-stream) has a similar structure as the C-stream, except the basic block consists of an MLP layer and a max pooling layer. Let the output of the C-stream and N-stream be F_C and F_N

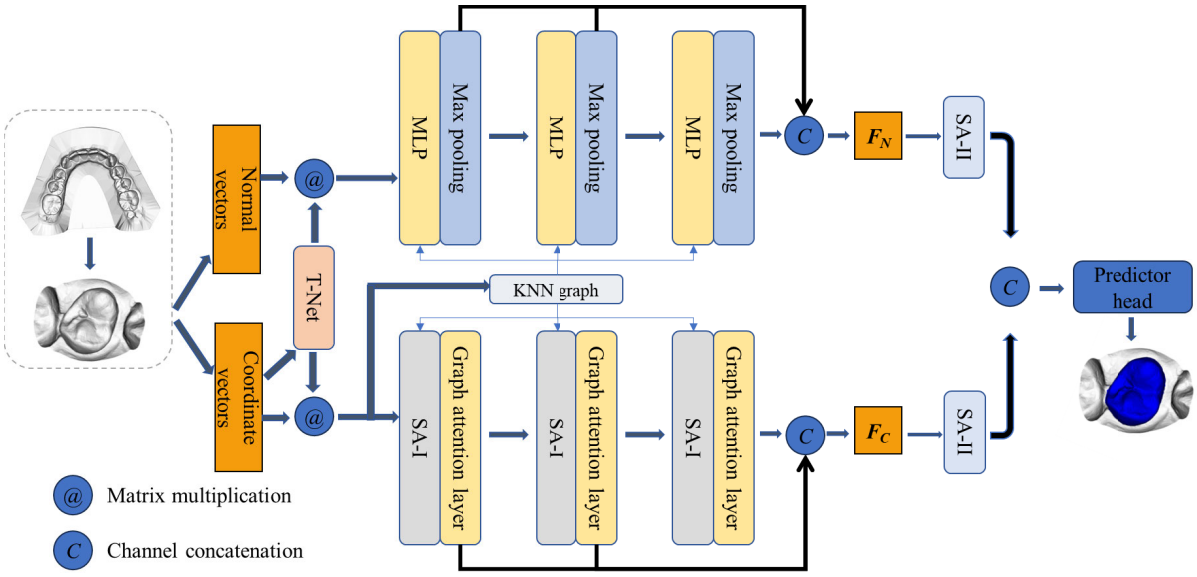


FIGURE 5. The architecture of TSGCN-SA.

respectively, they will go through an SA-II layer before the fusion of the two streams. The fused features will be used for the final prediction.

Within the above procedures, some issues need to be noted. The first one is that SA-I and SA-II are proposed for global feature extraction and adaptive feature fusion respectively, whose details are shown in Section III-B and Section III-C. Secondly, the ROI of the target tooth comes from the dental mesh. During the training phase, the ROIs are prepared ahead as a data set. During the testing or practical application, a user can use the proposed interface to extract them. Details of the proposed data set construction approach and user interaction are given in Section III-D. Last but not least, please refer to PointNet [14] for the details of T-Net, which is used for geometric alignment. The rest parts including the KNN graph and the predictor head are the same as in TSGCN [12].

B. SA-I BASED GLOBAL FEATURE EXTRACTION

A key point of TSGCN-SA is to combine the self-attention mechanism with graph convolution. As depicted in the coordinate stream in Fig. 5, before each graph attention layer, there is a self-attention layer named SA-I. The details of SA-I are shown in Fig. 6(a).

The inspiration for the SA-I layer comes from the Efficient Attention [20] and Point Transformer [14]. To reduce the computational complexity, the scalar attention scheme is adopted in SA-I, and the calculation order is adjusted according to [20]. To improve the accuracy, position encoding ζ is introduced into the attention calculation as done in [14].

C. SA-II BASED STREAM FUSION

Channel concatenation is a common practice to integrate multiple features. But before concatenating the outputs of the coordinate stream F_C and normal stream F_N, TSGCN points

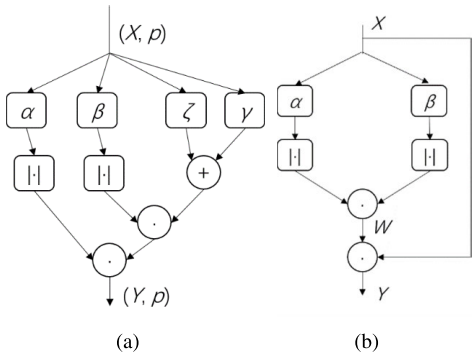


FIGURE 6. Two self-attention modules proposed in TSGCN-SA. (a) SA-I, (b) SA-II.

out that it is necessary to perform normalization to eliminate interference caused by scale differences. Specifically, let f_C^i and f_N^i represent the *ith* row of F_C and F_N respectively, then the normalization can be described by Eq. (3) and (4).

$$\hat{f}_C^i = \delta_C f_C^i = \frac{|f_N^i|}{|f_N^i| + |f_C^i|} f_C^i \tag{3}$$

$$\hat{f}_N^i = \delta_N f_N^i = \frac{|f_C^i|}{|f_N^i| + |f_C^i|} f_N^i \tag{4}$$

where δ_C and δ_N serve as normalization factors.

In this paper, we argue that in addition to the normalization, the contribution of each stream should also be considered during the fusion. Based on this theory, an attention layer named SA-II is introduced in TSGCN-SA to replace the aforementioned normalization process, and its operation details are shown in Fig. 6(b). The attention factor W is determined through learnable operations of α and β, which can simultaneously perform the role of normalization and contribution adjustment.

(a) (b)

FIGURE 7. The annotation used in our method. (a) A result of the proposed annotation transfer method, (b)The final annotation for individual tooth segmentation.

D. IMPLEMENTATION DETAILS

1) DATASET

The experimental data used in this work are originally from the MICCAI challenge dataset named 3DTeethSeg [15], which contains a total of 1,200 meshes including both upper and lower jaws. Each tooth on the jaw is labeled using the universal Federation Dentaire Internationale (FDI) annotation system.

To achieve the goal of deep learning based individual tooth segmentation, modifications are needed. The first modification is to cut the dental mesh into local parts which contain individual teeth automatically. To do that, a sphere is used. Let D, F, and P be the given dental mesh, the FDI number of the target tooth, and a set of vertexes of D who are labeled as F respectively. The sphere can be determined by the center c and the radius R, where c is the mean coordinates of P, and R is the max distance between c and any point in P. In practice, to improve the system robustness, we use αR instead of R to achieve the cutting, where $\alpha = 1.3$, 1.5, and 1.7 respectively.

The second modification is about the annotation, which is critical to supervised learning. Since our method uses the face-based scheme, each face should be labeled appropriately. However, the annotation of 3DTeethSeg data is based on vertexes. To solve this problem, a conservative strategy is presented to transfer the annotation from the vertex-basis to the face-basis. Let $L(p1)$, $L(p2)$ and $L(p3)$ be the label of three vertexes $p1$, $p2$ and $p3$ respectively, which defines a triangular face F in the given mesh. The label of F will be determined by Eq. (5). Fig. 7(a) illustrates a typical result of the proposed annotation transfer method.

$$L(F) = \begin{cases} L(p_1), & L(p_1) = L(p_2) = L(p_3) \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

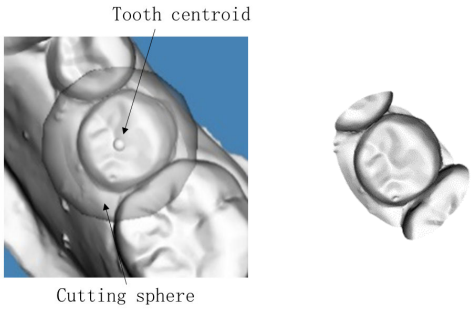
In addition to the transfer, a 4-class annotation system is used, namely T1, T2, T3, and T4 corresponding to the incisor, canine, premolar, and molar shown in Table 1. Adults normally have up to 28 teeth. The reason for reducing the number of classes from 28 to 4 is that teeth within each of the 4 classes are of similar shape, which makes it hard for a deep learning model based on shape features to tell. Finally, the mesh region that represents non-target teeth will be treated as background, even if it belongs to a tooth (see Fig. 7(b)).

TABLE 1. The relationship between 4-class names used in this paper and the FDI numbers.

Class name	FDI numbers	Tooth name
T1	11, 12, 21, 22, 31, 32, 41, 42	Incisor
T2	13, 23, 33, 43	Canine
T3	14, 15, 24, 25, 34, 35, 44, 45	Premolar
T4	16, 17, 18, 26, 27, 28, 36, 37, 38, 46, 47, 48	Molar

TABLE 2. A list of the amount and distribution of data for each class.

Class name	Number	Proportion (%)
T1	5184	27.88
T2	2809	15.11
T3	5700	30.67
T4	4896	26.34



(a) (b)

FIGURE 8. Demonstration of the user interface to fulfill the cutting. (a) Before the cutting, (b) After the cutting.

2) ALIGNMENT

The geometric differences in data, such as the position variances, may disturb the model prediction. That is a reason why a normalization process is needed before training and inference in general. Though data in 3DTeethSeg have been aligned to a uniform position, orientation, and scale already, the ROI data resulting from the above-cutting process are no longer aligned.

To deal with that, a T-Net mini-network is added (see Fig. 5), which can learn a transformation matrix to align the feature to a canonical space. Details of the T-Net architecture can be found in [21].

A key for T-Net based alignment is to increase the diversity of training data. Therefore, random translation, rotation, and scaling operations were applied for data augmentation. The final amount and distribution of data are shown in Table 2.

3) USER INTERFACE

During the testing, a convenient and intuitive user interface is involved to determine the ROI for each target tooth segmentation. As shown in Fig. 8(a), when the user's mouse moves near a tooth, a sphere will appear which illustrates the cutting area that can produce the ROI as shown in Fig. 8(b).

Hiding behind the user interface is a lightweight center point prediction network, which is pre-trained using the approach introduced in TSegNet [17]. The sphere candidates

TABLE 3. The statistical radius (R) of tooth based on 3DTeethSeg.

FDI	R(mm)	FDI	R(mm)	FDI	R(mm)	FDI	R(mm)
11	4.54	21	4.69	31	4.32	41	3.51
12	3.62	22	3.53	32	3.85	42	3.78
13	3.66	23	3.95	33	4.12	43	3.56
14	4.73	24	4.64	34	4.28	44	4.31
15	5.57	25	5.21	35	6.24	45	6.11
16	5.92	26	6.17	36	6.05	46	6.03
17	5.43	27	5.45	37	4.69	47	4.93

TABLE 4. A list of methods for comparison.

Name	Year of publication	Source	Type
PN++	2017	NIPS	Point-based
PT	2021	ICCV	Point-based
MeshSegNet	2020	TMI	Face-based
TSGCN	2022	TMI	Face-based

are centered at the predicted tooth centers, while the radiuses are preset according to the statistical values from 3DTeethSeg as shown in Table 3. When an abnormal tooth is encountered, the user can adjust the sphere by clicking and dragging (for position tuning) or scrolling (for radius tuning) the mouse.

IV. EXPERIMENT AND RESULT

Experiments are conducted to assess the proposed method. They were running on a desktop computer with an NVIDIA RTX 3090 (24GB) GPU. The Python 3.7, PyTorch 1.12, and CUDA 11.6 are used for programming.

For comparison, closely related methods including PointNet++ (PN++) [7], Point Transformer (PT) [21], MeshSegNet [8], and TSGCN [12] are selected. Among them, PN++ and PT are methods based on points, while MeshSegNet and TSGCN are based on faces. More details of the methods for comparison can be found in Table 4.

A. VISUALIZATION

First, we test the proposed method with multiple dental meshes. Among them, 4 typical results are visualized in Fig. 9 in rows. Different colors in the figure represent different categories ranging from T1 to T4. Columns from left to right are the GT, the results of PN++, PT, MeshSegNet, TSGCN, and ours respectively. It can be observed that lots of prediction errors occur by PN++ and MeshSegNet, including mistaking foregrounds as backgrounds (see Raw 2) and vice versa (see Raw 4), or mistaking a foreground class as a different foreground class (see Raw 3). By contrast, PT performs slightly better own to the Transformers. However, it is inferior to TSGCN because TSGCN is trained with richer features and a multi-branch architecture. Thanks to the proposed SA modules, our method outperforms the others even for the challenge case shown in the first raw.

To evaluate the performance quantitatively, the overall accuracy (OA) and the mean Intersection-over-Union (mIoU) are selected as metrics. OA represents the percentage of the number of correctly predicted faces over the total number of faces. Let $IoU(c)$ be the intersection-over-union of a class c that belongs to the set of categories $\mathbb{C}$, then $mIoU$ can be calculated as Eq. (6):

$$mIoU = \frac{\sum_{c\in\mathbb{C}} \mathrm{IoU}(c)}{|\mathbb{C}|} \tag{6}$$

where $|\mathbb{C}|$ means the number of $\mathbb{C}$.

The comparison results in terms of OA and mIoU are shown in Fig. 10 and Fig. 11 respectively. Within each figure, there are both results focused on each category and results overall. As the overall results show, our method can achieve the best performance among the comparisons in terms of all metrics.

TABLE 5. The notation of methods in the ablation study.

	With SA-I	Without SA-I
With SA-II	TSGCN-SA	Ours(Conv, SA-II)
Without SA-II	Ours(SA-I, Nor)	TSGCN

TABLE 6. The results of the ablation study.

Method	OA(%)	mIoU(%)
TSGCN-SA	97.53	94.5
Ours(SA-I, Nor)	97.12	94.23
Ours(Conv, SA-II)	96.83	94.14
TSGCN	96.45	93.55

TABLE 7. The number of parameters and average time for inference of the proposed method.

Method	Number of parameters(MB)	Average time for inference (MS)
Ours	6.12	135
TSGCN	3.76	68

B. ABLATION STUDY

Compared to the baseline model (i.e., TSGCN), core improvements of TSGCN-SA include the usage of SA-I and SA-II to replace the convolution (Conv) layers and the normalization (Nor) layers in TSGCN respectively. To investigate the impact of SA-I and SA-II modules, ablation studies are carried out among methods with or without SA-I/II, whose notations are shown in Table 5.

The results of the ablation study are recorded in Table 6. It can be noticed that the solo usage of SA-I and SA-II results in 0.67 and 0.38 improvement in terms of OA, 0.68 and 0.59 improvement in terms of mIoU respectively. It means that SA-I contributes more than SA-II. Anyway, the replacements are beneficial to the system's performance.

Next, the proposed method is evaluated in terms of the number of parameters and average time for inference. As shown in Table 7, they equal 6.12 megabytes (MB) and 135 milliseconds (MS) respectively for a mesh of ten thousand faces. Compared to TSGCN, our model possesses more parameters, which means a larger capability to solve problems in deep learning in general. On the other

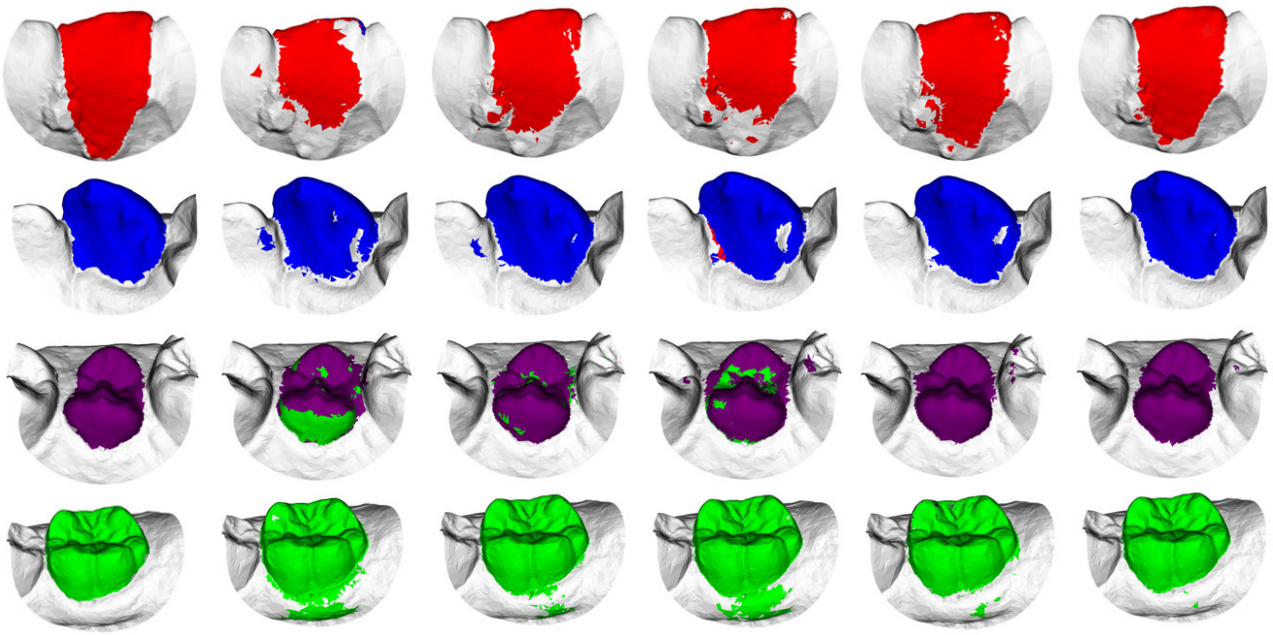


FIGURE 9. Visualization of the comparison results. (a) GT, (b) PN++, (c) PT, (d) MeshSegNet, (e) TSGCN, (f) Ours.

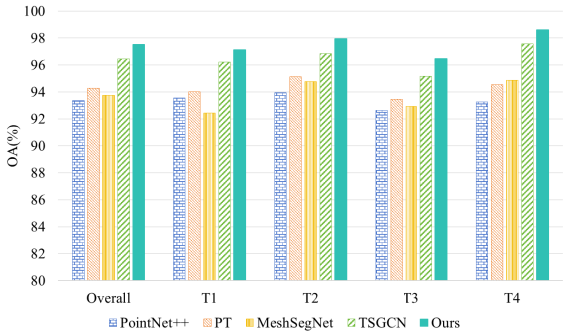


FIGURE 10. Comparison results in terms of OA.

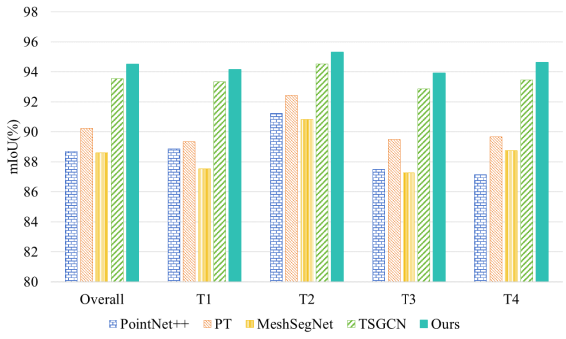


FIGURE 11. Comparison results in terms of mIoU.

hand, our method takes around 0.1 seconds to segment one tooth, which meets the standard for real-time user feedback.

TABLE 8. The results of the generalization experiment.

Method	OA(%)	mIoU(%)
PN++	85.31	79.28
PT	85.47	79.63
MeshSegNet	83.25	78.36
TSGCN	86.48	80.18
Ours	91.44	85.62

C. GENERALIZATION EXPERIMENT

To verify the scalability and generalization ability of the proposed method, experiments using a new dataset with over 1,400 teeth which constructed from private commercial dental data are carried out for inference without further training.

The results are recorded in Table 8, which shows an inferior outcome for each method working on the new data than itself working on 3DTeethSeg. This phenomenon is quite common because the two datasets must have different data distributions. However, our method still outperforms the others in this situation. Furthermore, few rounds of training with new data will relieve this problem and boost the performance normally in our experience.

V. LIMITATIONS AND FUTURE WORK

Despite the strengths of the proposed method, there are still some limitations. For example, there is an imbalance per label as illustrated in Table 2, due to the 4-class annotation system used in our method. As we mentioned before, though there are up to 28 teeth in the upper and lower jaws, the 4-class strategy makes it easier for a deep learning model to distinguish the

differences between each class, thus resulting in a better performance. To deal with the imbalance issue, the mesh simplification method and sampling method could be used in future work. Another issue is the identification of a target tooth with its FDI number. Though the model we trained is not able to achieve that, it can easily be done because we have a lightweight tooth centroid prediction network, which is supervised with the FDI labels. However, training a model that can predict 28 classes is another research direction in the future, especially for a segmentation method that can segment every tooth from the dental mesh at once automatically.

VI. CONCLUSION

A deep learning method named TSGCN-SA is proposed in this paper to segment individual teeth from dental mesh. Compared to the baseline model TSGCN, which is one of the state-of-the-art, TSGCN-SA is superior in the global understanding capability because of the proposed self-attention modules SA-I and SA-II, which can facilitate effective feature extraction and stream fusion. In addition, an automatic training data generation scheme is proposed to ensure supervised learning from big data during the training phase, and an intuitive user interface is involved for the intuitive and fast segmentation during the application phase.

A large number of experimental results show that TSGCN-SA outperforms the SOTAs in terms of overall accuracy and mean Intersection-over-Union. With the help of effective cooperation between global and local features, the proposed method is competent in the task of individual tooth mesh segmentation.

ACKNOWLEDGMENT

The authors gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] S. Tian, N. Dai, B. Zhang, F. Yuan, Q. Yu, and X. Cheng, "Automatic classification and segmentation of teeth on 3D dental model using hierarchical deep learning networks," *IEEE Access*, vol. 7, pp. 84817–84828, 2019.

[2] M. Hashem, M. L. Mohammed, and A. E. Youssef, "Improving the efficiency of dental implantation process using guided local search models and continuous time neural networks with robotic assistance," *IEEE Access*, vol. 8, pp. 202755–202764, 2020.

[3] G. Badiali, L. Cercenelli, S. Battaglia, E. Marcelli, C. Marchetti, V. Ferrari, and F. Cutolo, "Review on augmented reality in oral and cranio-maxillofacial surgery: Toward 'surgery-specific' head-up displays," *IEEE Access*, vol. 8, pp. 59015–59028, 2020.

[4] B.-J. Zou, S.-J. Liu, S.-H. Liao, X. Ding, and Y. Liang, "Interactive tooth partition of dental mesh base on tooth-target harmonic field," *Comput. Biol. Med.*, vol. 56, pp. 132–144, Jan. 2015.

[5] S.-H. Liao, S.-J. Liu, B.-J. Zou, X. Ding, Y. Liang, and J.-H. Huang, "Automatic tooth segmentation of dental mesh based on harmonic fields," *BioMed Res. Int.*, vol. 2015, pp. 1–10, Jul. 2015.

[6] R. Hanocka, A. Hertz, N. Fish, R. Giryes, S. Fleishman, and D. Cohen-Or, "MeshCNN: A network with an edge," *ACM Trans. Graph.*, vol. 38, no. 4, pp. 1–12, Aug. 2019.

[7] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: Deep hierarchical feature learning on point sets in a metric space," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, Long Beach, CA, USA, 2017, pp. 5105–5114.

[8] C. Lian, L. Wang, T.-H. Wu, F. Wang, P.-T. Yap, C.-C. Ko, and D. Shen, "Deep multi-scale mesh feature learning for automated labeling of raw dental surfaces from 3D intraoral scanners," *IEEE Trans. Med. Imag.*, vol. 39, no. 7, pp. 2440–2450, Jul. 2020.

[9] Z. Wang, J. Na, and B. Zheng, "An improved kNN classifier for epilepsy diagnosis," *IEEE Access*, vol. 8, pp. 100022–100030, 2020.

[10] X. Xu, C. Liu, and Y. Zheng, "3D tooth segmentation and labeling using deep convolutional neural networks," *IEEE Trans. Vis. Comput. Graphics*, vol. 25, no. 7, pp. 2336–2348, Jul. 2019, doi: 10.1109/TVCG.2018.2839685.

[11] Y. Zhao, L. Zhang, C. Yang, Y. Tan, Y. Liu, P. Li, T. Huang, and C. Gao, "3D dental model segmentation with graph attentional convolution network," *Pattern Recognit. Lett.*, vol. 152, pp. 79–85, Dec. 2021.

[12] Y. Zhao, L. Zhang, Y. Liu, D. Meng, Z. Cui, C. Gao, X. Gao, C. Lian, and D. Shen, "Two-stream graph convolutional network for intra-oral scanner image segmentation," *IEEE Trans. Med. Imag.*, vol. 41, no. 4, pp. 826–835, Apr. 2022, doi: 10.1109/TMI.2021.3124217.

[13] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, Oct. 2021, pp. 9992–10002.

[14] H. Zhao, L. Jiang, J. Jia, P. Torr, and V. Koltun, "Point transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Montreal, QC, Canada, 2021, pp. 16259–16268.

[15] A. Ben-Hamadou, O. Smaoui, and A. Rekik, "3DTeethSeg'22: 3D teeth scan segmentation and labeling challenge," 2023, *arXiv:2305.18277*.

[16] M. Li and L. Nan, "Feature-preserving 3D mesh simplification for urban buildings," *ISPRS J. Photogramm. Remote Sens.*, vol. 173, pp. 135–150, Mar. 2021.

[17] Z. Cui, C. Li, N. Chen, G. Wei, R. Chen, Y. Zhou, D. Shen, and W. Wang, "TSegNet: An efficient and accurate tooth segmentation network on 3D dental model," *Med. Image Anal.*, vol. 69, Apr. 2021, Art. no. 101949, doi: 10.1016/J.MEDIA.2020.101949.

[18] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, Long Beach, CA, USA, 2017, pp. 6000–6010.

[19] H. Zhao, J. Jia, and V. Koltun, "Exploring self-attention for image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 10073–10082, doi: 10.1109/CVPR42600.2020.01009.

[20] S. Zhuoran, Z. Mingyuan, Z. Haiyu, Y. Shuai, and L. Hongsheng, "Efficient attention: Attention with linear complexities," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Waikoloa, HI, USA, Jan. 2021, pp. 3530–3538, doi: 10.1109/WACV48630.2021.00357.

[21] R. Q. Charles, H. Su, M. Kaichun, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 77–85, doi: 10.1109/CVPR.2017.16.

SHI-JIAN LIU received the B.S. degree in mathematics from Xiangtan University, Xiangtan, China, in 2006, the M.S. degree in computer science from Changsha University of Science and Technology, Changsha, China, in 2010, and the Ph.D. degree in computer science from Central South University, Changsha, in 2015. In 2018, he was a Visiting Scholar with Manchester Institute of Biotechnology, The University of Manchester, Manchester, U.K. He is currently an Associate Professor with the School of Computer Science and Mathematics, Fujian University of Technology, Fuzhou, China. His research interests include computer vision, deep learning, and computer graphics.

CHAO-MING KANG received the B.S. degree in information security from Fujian Police College, Fuzhou, China, in 2021. He is currently pursuing the master's degree with Fujian University of Technology. His research interests include mesh processing and deep learning.

FENG-HUA HUANG (Member, IEEE) received the Ph.D. degree in geographic information systems from Fujian Normal University. He is currently the Dean of the Academy of Intelligent Engineering Technology and the Vice Dean of the College of Artificial Intelligence, Yango University, China. He is also the Director of Fujian Key Laboratory of Spatial Information Perception and Intelligent Processing and Fujian University Engineering Research Center of Spatial Data Mining and Application. His research interests include big data analysis, pattern recognition, and remote sensing images procession.

ZHENG ZOU received the B.S. degree in mathematics and the M.S. degree in computer science from Changsha University of Science and Technology, Changsha, China, in 2006 and 2010, respectively, and the Ph.D. degree in computer science from Central South University, Changsha, in 2017. She is currently a Lecturer with the College of Computer and Cyber Security, Fujian Normal University, Fuzhou, China. Her research interests include biomedical image processing and deep learning.