Weakly Supervised Tooth Instance Segmentation on 3D Dental Models with Multi-Label Learning

Haoyu Wang[1,+], Kehan Li[3,+], Jihua Zhu[1,(✉)], Fan Wang[3], Chunfeng Lian[2,4,(✉)], and Jianhua Ma[3,4,(✉)]

[1] School of Software Engineering, Xian Jiaotong University, Xi'an, China
zhujh@xjtu.edu.cn
[2] School of Mathematics and Statistics, Xian Jiaotong University, Xi'an, China
chunfeng.lian@xjtu.edu.cn
[3] Key Laboratory of Biomedical Information Engineering of Ministry of Education, School of Life Science and Technology, Xian Jiaotong University, Xi'an, China
{fan.wang,jhma}@xjtu.edu.cn
[4] Pazhou Lab(Huangpu), Guangzhou, China

Abstract. Automatic tooth segmentation on 3D dental models is a fundamental task for computer-aided orthodontic treatment. Many deep learning methods aimed at precise tooth segmentation currently require meticulous point-wise annotations, which are extremely time-consuming and labor-intensive. To address this issue, we propose a **w**eakly **s**upervised **t**ooth **i**nstance **s**egmentation network (WS-TIS) with multi-label learning, which only requires subject-level class labels along with approximately 50% of point-wise tooth annotations. Our WS-TIS consists of two stages, including fine-grained multi-label classification and tooth instance segmentation. Precise tooth localization is frequently pivotal in instance segmentation. However, annotation of tooth centroids or bounding boxes is often challenging when we have limited point-wise tooth annotations. Therefore, we design a proxy task to weakly supervise tooth localization. Specifically, we utilize a fine-grained multi-label classification task, equipping with the disentangled re-sampling strategy and a gated-attention mechanism, which can assist the network in learning discriminative tooth features. Based on discriminative features, discriminative regions can be easily obtained, thereby accurately cropping each tooth. In the second stage, a segmentation module is trained on limited annotated data (approximately 50% of all teeth) to accurately segment each tooth within the cropped regions. Experiments on Teeth3DS demonstrate that our WS-TIS achieves superior performance compared to state-of-the-art approaches under full annotations. Our code will be released on https://github.com/ladderlab-xjtu/WS-TIS.

Keywords: Instance segmentation · Weak supervision · Limited annotations · Multi-label discriminative localization · 3D dental models

[+] The authors contribute equally to this work.

1 Introduction

The 3D dental models acquired by intraoral scanners (IOS) are becoming increasingly popular in clinical computer-aided design (CAD) due to their convenience and radiation-free nature. As a very crucial step in clinical orthodontics, precise tooth segmentation serves as the foundation for subsequent tooth realignment and treatment planning. However, precise point-wise manual annotation is time-consuming and labor-intensive, which has prompted many researchers to focus on developing methods for achieving accurate fully automatic tooth segmentation. For example, many handcraft geometric feature-based methods [6,17] and CNN-based methods [21,15] have been proposed. However, these methods either suffer from the reliance on manual interaction, poor robustness, or the limitation of insufficient resolution due to excessive memory consumption, resulting in inaccurate segmentation results.

In the community of computer vision, many pioneering works [10,16,20,11] have been proposed to consume raw point cloud data without any data format transformation, thus avoiding information loss. Inspired by them, many tooth segmentation approaches capable of directly handling raw point cloud data have been proposed [7,18,22,4,12]. For example, MeshSegNet [7] improves PointNet by enabling the network to learn local relationships between points, enhancing the accuracy of fine-grained tooth segmentation tasks. TSGCNet [22] learns different raw geometric information (i.e., coordinates and normal vectors) through two different streams, eliminating isolated false predictions caused by the confusion between different geometric features. Differing from the semantic segmentation methods mentioned above, TSegNet [4] performs tooth delineation in the instance segmentation fashion. Specifically, TSegNet first predicts the centroid of each tooth through a centroid regression network. Then, a segmentation network is trained to segment each tooth from the cropped areas obtained by the centroid points. However, these existing methods require dense point-wise annotations, which are labor-intensive for high-resolution dental models. The difficulty of obtaining such a large amount of dense point-wise annotations in reality also hinders the generalization and practical application of these methods. To reduce the annotation data, DArch [12] proposes a weakly annotated training approach to train the segmentation network using only a few annotation teeth from each dental model. However, the first stage of DArch still requires complete centroid annotations, which is challenging to obtain in the absence of dense point-wise annotations.

To address this issue, we propose a weakly supervised tooth instance segmentation network (WS-TIS) with multi-label learning to accurately segment each tooth with limited annotation data. To accurately locate teeth without tooth centroid annotations, we leverage fine-grained classification as a proxy task to learn discriminative features with multi-label learning. In our proxy task, we use a task-oriented re-sampling strategy named disentangled re-sampling to alleviate the feature entanglement caused by extreme label co-occurrence. To enhance the discrimination of feature representations, we introduce a gated-attention mechanism, allowing features from different channels to focus on each individual tooth.

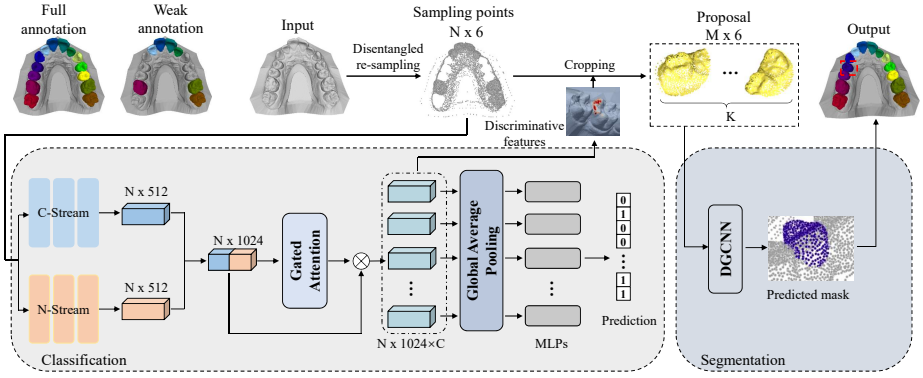

Fig. 1. The schematic diagram of our WS-TIS network.

In this way, our network can obtain accurate classification results and effectively learn the discriminative representations of each tooth with multi-label learning. Based on discriminative features, weakly supervised localization techniques (e.g., CAM [24] and Grad-CAM [14]) can be easily employed to localize the position of each tooth. Then, the localized teeth are cropped to train a binary segmentation network with weak point-wise annotations. Extensive experiments have shown that our network can accurately learn the discriminative features of each tooth under weak supervision. Besides, the segmentation performance of our WS-TIS under weak annotations is superior even compared with the state-of-the-art tooth segmentation methods under full annotations.

2 Method

2.1 Overview

As shown in Fig. 1, our WS-TIS can be divided into two parts, i.e., tooth classification network and tooth instance segmentation network. The input of the first network are coordinates and normal vectors, which can be denoted as an N×6 matrix, with N standing for the number of points in the dental model, 6-dimensional standing for 3-dimensional coordinates and 3-dimensional normal vectors. Considering the superior performance of TSGCNet in tooth semantic segmentation, we select it as the backbone for the feature extraction. Equipped with the disentangled re-sampling strategy and gated-attention mechanism, our network can accurately predict multi-label classification results and learn the discriminative features of each tooth. By leveraging the discriminative feature, the feature visualization techniques can easily detect the position of each tooth. The regions containing the target tooth are cropped and fed into the instance segmentation network, which can be denoted as an M×6 matrix, with M standing for the number of points in the cropped region and 6-dimensional still standing for coordinates and normal vectors. The segmentation network can delineate each tooth within the cropped area.

2.2 Tooth multi-label classification

In practice, obtaining the centroids or bounding boxes of the target objects is challenging. To accurately localize each tooth, we leverage a fine-grained classification in a weakly supervised manner. Specifically, we train a multi-label classification network as a proxy task to detect the discriminative regions of each tooth. Leveraging two distinct feature extraction streams of TSGCNet, i.e., C-Stream and N-Stream, the learned geometric representations will be further fused to make the final prediction. To make accurate classification and enable the network to learn discriminative features, we design a disentangled re-sampling strategy to alleviate the feature entanglement caused by extreme label co-occurrence and further introduce a gated-attention mechanism to enhance the feature discrimination of different teeth. We will provide detailed explanations in the following sections.

Disentangled re-sampling strategy In the fine-grained classification task, discriminative features typically indicate the position of the target category. However, due to the label co-occurrence, certain labels often appear together in multi-label classification tasks, thereby affecting the learning of discriminative features. Furthermore, the imbalanced data distribution further increases the difficulty of learning. Many multi-label classification networks [19,3,13] are dedicated to addressing this issue by model and loss designs. However, unlike natural images (for example, toothbrushes always appear with people), label co-occurrence in dental surfaces is not explicitly linked. In addition, the label co-occurrence issue in dental models is particularly severe since most people have a full set of teeth, resulting in a scarcity of negative samples in the dataset. This leads to feature entanglement between different teeth, making it difficult for the network to learn discriminative features. Since the aforementioned issues, most existing methods are not suitable for our task.

Thus, we propose a task-oriented re-sampling strategy (not required during inference), i.e., a disentangled re-sampling strategy. Different from common multi-label classification tasks, our classification task serves as a proxy task for tooth localization before instance segmentation, which means that we have partial tooth point-wise labels. Specifically, we randomly select K teeth that have point-wise labels on a dental model and put them into the classification network. The input can be denoted as a $P\times6$ matrix. This simple but intuitive design effectively balances the number of instances between different classes and significantly alleviates the issue caused by extreme label co-occurrence, allowing our network to learn the discriminative features of different teeth.

Gated-attention mechanism Furthermore, discriminative features should be as accurate as possible, ensuring they are contained within the range of the target teeth. To make the features more discriminative for different tooth categories, we hope to introduce an attention mask to weight features. Inspired by [5], we introduce a gated-attention mechanism in our classification network. Specifically, after the feature fusion of two different streams, we obtain a multi-view high-level representation which can be denoted as $\mathbf{f} \in \mathbb{R}^{N\times1024}$, where N is the number

of points and 1024 is the number of channels. The attention mask $\mathbf{A}$ can be obtained from multi-view representation $\mathbf{f}$, which can be denoted as

$$\mathbf{A} = Softmax(\omega(tanh(\phi_{\theta_1}(\mathbf{f_i})) \odot sigmoid((\phi_{\theta_2}(\mathbf{f_i}))), \tag{1}$$

where ω, θ_1 and θ_2 are learnable parameters, *tanh* and *sigmoid* are activation function. It is worth noting that $\mathbf{A} \in \mathbb{R}^{N \times C}$ represents the relevance of each point to the different tooth classes, where C denotes the number of teeth in dental models. The learned attention mask $\mathbf{A}$ will be used to weight the multi-view feature $\mathbf{f}$, which can be denoted as:

$$\hat{\mathbf{f}} = \mathbf{A} \times \mathbf{f} \tag{2}$$

where $\hat{\mathbf{f}} \in \mathbb{R}^{N \times 1024 \times C}$. The proposition of attention mask $\mathbf{A}$ allows the high-level representation $\mathbf{f}$ to exhibit explicit discriminative characteristics for different teeth. Thus, localization techniques (e.g., CAM and Grad-CAM) can be employed to accurately detect the position of all teeth in the setting of multi-label learning. Based on $\hat{\mathbf{f}}$ we perform binary classifications separately for each feature which represents different tooth categories to determine if this tooth exists on the dental model. To further alleviate the impact of class imbalance issues, we employ focal loss[8] in our multi-label classification network.

2.3 Tooth instance segmentation

To ensure the complete cropping of the target tooth, we select the nearest M points to the discriminative features' centroid as the cropped region, which is set as $2,048$ in our experiments. It is worth noting that we only crop teeth with point-wise labels on a dental model. The cropped region which can be denoted as an $M \times 6$ matrix will be put into the segmentation network for binary segmentation. Considering the effectiveness and convenience, we choose the DGCNN as our segmentation network. It will ultimately delineate the target teeth within the cropped region. In the inference stage, we fuse the segmentation results of each tooth onto a complete dental model based on the multi-label classification results. It is worth noting that, we already have the IDs for each cropped region in the classification network, so there is no need to design an additional tooth ID prediction network as in TSegNet [4].

2.4 Implementation details

Our WS-TIS is trained on an NVIDIA RTX4090 GPU. We employ focal loss for the multi-label classification network and cross-entropy loss for the instance segmentation network. The learning rate for both networks is set as 0.001, and they are trained for 200 epochs. The mini-batch sizes of the two networks are 2 and 32, respectively. Besides, we sample $8,000$ points as input to the classification network and crop $2,048$ points for the segmentation network.

3 Experiments

3.1 Dataset and evaluation metrics

To evaluate the performance of our model, we use a publicly available dental dataset, Teeth3DS [2]. The Teeth3DS releases $1,200$ high-resolution dental

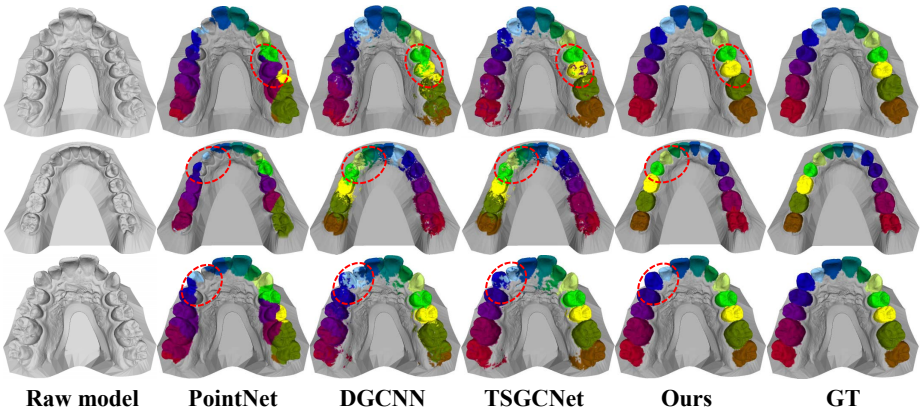

Fig. 2. The visualization results of different segmentation methods. The red circles highlight the locations where our method outperforms others significantly.

Table 1. The classification and segmentation results (mean±std) of our method under weak annotations and other competing methods under full annotations.

Method	Classification		Segmentation	
	mAP ↑	FPR↓	mIoU ↑	Dice↑
PointNet [10]	-	-	79.84 ± 2.74	82.43 ± 1.46
DGCNN [16]	-	-	84.04 ± 2.47	87.04 ± 1.22
TSGCNet [22]	-	-	90.46 ± 2.13	91.61 ± 1.56
DB Loss [19]	94.70 ± 1.00	46.71 ± 2.31	-	-
WS-TIS	$\mathbf{96.37 \pm 0.74}$	$\mathbf{17.50 \pm 1.19}$	$\mathbf{96.09 \pm 0.94}$	$\mathbf{97.37 \pm 0.48}$

models collected from intraoral scanners, which have been annotated with dense point-wise labels. The number of points in the dental models from Teeth3DS ranges from $100,000$ to $400,000$. Due to limitations in computational resources, we down-sample these points to approximately $16,000$. Notably, to simulate the limited label scenario, we randomly mask some teeth as background.

We randomly select 600 samples for training, 300 samples for validation, and the remaining 300 samples for testing. In the comparison experiments, we utilize the commonly used mIoU and Dice metrics to quantify the performance of our segmentation network. Due to the extreme label co-occurrence in the dental models, we employ mAP (mean Average Precision) and FPR=FP/(FP+TN) (False Positive Rate) to quantify the performance of our classification network.

3.2 Comparison results

Competing methods To validate the performance of our method, we compare our WS-TIS with the state-of-the-art methods of tooth segmentation and multi-label classification. For the multi-label classification, we compare our method with **DB Loss** [19] which can effectively address the issues caused by label co-occurrence and imbalanced data distribution in natural images. For the tooth segmentation task, we compare our method with **TSGCNet** [22], **PointNet** [10], and **DGCNN** [16]. TSGCNet utilizes two distinct streams to learn the high-level

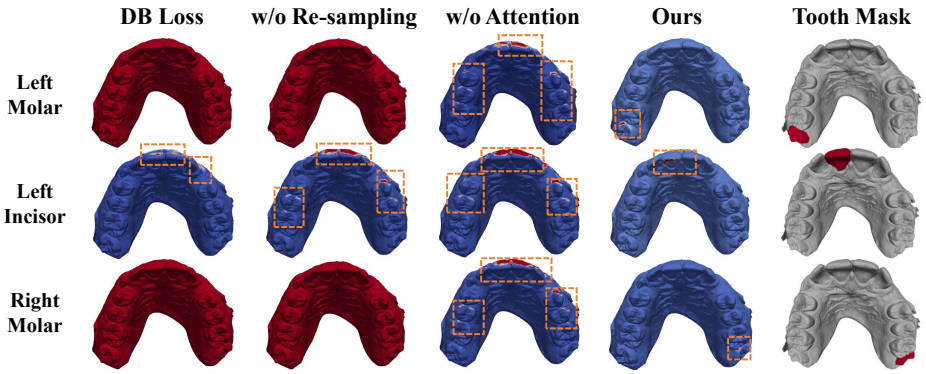

Fig. 3. Feature visualization of classification network with different experimental settings. The blue indicates regions with low feature values, while the red highlights areas with higher feature values, marked by dashed squares. The last row displays the positions of the target teeth.

representations of different geometric attributes, obtaining superior performance in tooth segmentation. PointNet and DGCNN are representative segmentation networks in general 3D shape segmentation.

Results The quantitative results of multi-label classification are shown in Table 1. From Table 1, we can observe that the classification network of our WS-TIS achieves superior performance on both mAP and FPR. While our method only has a slight lead of 1.6% in mAP, it achieves a nearly 30% reduction in FPR compared to DB loss. This demonstrates the excellent performance of our method in multi-label tooth classification, especially in cases of extreme label co-occurrence. We further employ feature visualization techniques (we choose CAM for its convenience and simplicity) to show the locations of learned discriminative features, as illustrated in Fig. 3. Obviously, our WS-TIS learns discriminative features that can accurately localize each tooth, while DB Loss in the first row struggles to differentiate between individual teeth.

For the tooth segmentation, the quantitative results are shown in Table 1. From Table 1, we can observe that our network trained with weak annotations outperforms other segmentation networks under full-label supervision. Compared to the TSGCNet, our method achieves an improvement of 5.6% in mIoU and 5.8% in Dice. Fig. 2 presents the visualization results of different segmentation networks. From Fig. 2, we can observe that due to the proximity and similar appearances of adjacent teeth, other methods often make false predictions in these regions. However, the high-quality discriminative features make our method consistently achieve stable and accurate segmentations.

We set up the comparison experiments with the same configuration to ensure a fair comparison. Notably, we do not include recent methods [1,9,23] in our comparative experiments as they don't release their codes. However, according to their reported results on the same dataset (Teeth3DS), our method remains qualitatively superior.

Table 2. Ablation studies regarding the key components of our method.

Method	mAP ↑	FPR ↓
w/o re-sampling	94.51	49.24
w/o attention	87.31	62.36
WS-TIS	**96.37**	**17.50**

3.3 Ablation studies

Sampling strategy To alleviate the feature entanglement caused by extreme label co-occurrence, we design a disentangled re-sampling strategy that effectively increases the number of negative samples, allowing the classification network to learn the discriminative features of each tooth. To validate the effectiveness of our disentangled re-sampling strategy, we conduct an ablation experiment without this task-oriented re-sampling strategy (w/o re-sampling).

According to the quantitative results shown in Table 2, we can observe that the method without the re-sampling strategy, despite achieving a high mAP, exhibits nearly 50% FPR. This indicates that networks struggle to learn discriminative features effectively in situations of extreme label co-occurrence, resulting in inaccurate classifications. The feature visualization results in Fig. 3 also confirm this claim. From Fig. 3, we can see that the network without sampling strategy tends to confuse the features of different teeth. On the contrary, our approach can accurately learn the discriminative features of each tooth, providing effective assurance for tooth localization.

Gated-attention We introduce a gated-attention mechanism to enhance the discriminative features among different teeth, thereby improving classification accuracy. To check the efficacy of our gated-attention mechanism, we conduct an ablation study by removing gated-attention from our WS-TIS.

The classification results are presented in Table 2. From it, we can observe that the gated-attention significantly improved classification accuracy, increasing mAP by 9% and reducing FPR by 45%. Furthermore, from the feature visualization results in Fig. 3, we observe that the network struggles to learn discriminative features for each tooth without gated-attention. This indicates that our gated-attention is effective in enhancing the discrimination of features among different teeth.

4 Conclusion

In this paper, we have proposed a weakly supervised tooth instance segmentation with multi-label learning to address tooth segmentation under weak annotations. Due to the absence of tooth center annotations, we adopt a fine-grained multi-label classification as a proxy task. Specifically, we design a disentangled re-sampling strategy to effectively alleviate the feature entanglement caused by extreme label co-occurrence, and leverage the gated-attention mechanism to enhance the discrimination of features among different teeth and thereby improve the accuracy of discriminative features. A cropped region can be obtained based

on the discriminative feature from the fine-grained classification network. Then, a segmentation network predicts the positions of the target tooth within the cropped area. Extensive comparison experiments demonstrate that our WS-TIS achieves superior performance in tooth segmentation even under weak annotations.

Acknowledgments. This work was supported in part by NSFC Grant (No. 62101430).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Almalki, A., Latecki, L.J.: Self-supervised learning with masked autoencoders for teeth segmentation from intra-oral 3d scans. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 7820–7830 (2024)
2. Ben-Hamadou, A., Smaoui, O., Chaabouni-Chouayakh, H., Rekik, A., Pujades, S., Boyer, E., Strippoli, J., Thollot, A., Setbon, H., Trosset, C., et al.: Teeth3ds: a benchmark for teeth segmentation and labeling from intra-oral 3d scans. arXiv preprint arXiv:2210.06094 (2022)
3. Cole, E., Mac Aodha, O., Lorieul, T., Perona, P., Morris, D., Jojic, N.: Multi-label learning from single positive labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 933–942 (2021)
4. Cui, Z., Li, C., Chen, N., Wei, G., Chen, R., Zhou, Y., Shen, D., Wang, W.: Tsegnet: An efficient and accurate tooth segmentation network on 3d dental model. Medical Image Analysis **69**, 101949 (2021)
5. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International Conference on Machine Learning. pp. 2127–2136. PMLR (2018)
6. Kumar, Y., Janardan, R., Larson, B., Moon, J.: Improved segmentation of teeth in dental models. Computer-Aided Design and Applications **8**(2), 211–224 (2011)
7. Lian, C., Wang, L., Wu, T.H., Wang, F., Yap, P.T., Ko, C.C., Shen, D.: Deep multi-scale mesh feature learning for automated labeling of raw dental surfaces from 3d intraoral scanners. IEEE Transactions on Medical Imaging **39**(7), 2440–2450 (2020)
8. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2980–2988 (2017)
9. Lucas Krenmayr, R.v.S., Daniel Schautd, P.R., Hafner, A.: Dilatedtoothsegnet: Tooth segmentation network on 3d dental meshes through increasing receptive vision. Journal of Imaging Informatics in Medicine (2024)
10. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 652–660 (2017)
11. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Advances in Neural Information Processing Systems **30** (2017)
12. 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)

13. Ridnik, T., Ben-Baruch, E., Zamir, N., Noy, A., Friedman, I., Protter, M., Zelnik-Manor, L.: Asymmetric loss for multi-label classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 82–91 (2021)
14. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 618–626 (2017)
15. 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)
16. Wang, Y., Sun, Y., Liu, Z., Sarma, S.E., Bronstein, M.M., Solomon, J.M.: Dynamic graph cnn for learning on point clouds. ACM Transactions on Graphics (tog) **38**(5), 1–12 (2019)
17. Wu, K., Chen, L., Li, J., Zhou, Y.: Tooth segmentation on dental meshes using morphologic skeleton. Computers & Graphics **38**, 199–211 (2014)
18. Wu, T.H., Lian, C., Lee, S., Pastewait, M., Piers, C., Liu, J., Wang, F., Wang, L., Chiu, C.Y., Wang, W., et al.: Two-stage mesh deep learning for automated tooth segmentation and landmark localization on 3d intraoral scans. IEEE Transactions on Medical Imaging **41**(11), 3158–3166 (2022)
19. Wu, T., Huang, Q., Liu, Z., Wang, Y., Lin, D.: Distribution-balanced loss for multi-label classification in long-tailed datasets. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16. pp. 162–178. Springer (2020)
20. Wu, W., Qi, Z., Fuxin, L.: Pointconv: Deep convolutional networks on 3d point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9621–9630 (2019)
21. Xu, X., Liu, C., Zheng, Y.: 3d tooth segmentation and labeling using deep convolutional neural networks. IEEE Transactions on Visualization and Computer Graphics **25**(7), 2336–2348 (2018)
22. Zhang, L., Zhao, Y., Meng, D., Cui, Z., Gao, C., Gao, X., Lian, C., Shen, D.: 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)
23. Zhijie Lin, Z.H., Xu Wang, B.Z., Chang Liu, W.S., Ji Tan, S.X.: Dbganet: dual-branch geometric attention network for accurate 3d tooth segmentation. IEEE Transactions on Circuits and Systems for Video Technology (2023)
24. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2921–2929 (2016)