Accurate 3D Tooth Segmentation via Skeleton and Centroid Guidance in CBCT Imaging

Muhammad Asif Jamal
Dept. of Information and Communication Engineering,
Gwangju, Republic of Korea.
asif.jamal@chosun.ac.kr

Bumshik Lee
Dept. of Information and Communication Engineering,
Gwangju, Republic of Korea.
bslee@chosun.ac.kr

***Abstract*—Accurate three-dimensional tooth segmentation from cone-beam computed tomography (CBCT) is critical for advanced dental diagnosis and treatment planning. This study presents a deep-learning approach for segmenting individual teeth from manually cropped teeth regions of interest (ROIs) in preprocessed CBCT images. The proposed method integrates dilated attention and squeeze-and-excitation attention mechanisms into the skip connections of a V-Net architecture, enabling precise segmentation of complex dental structures, including tooth roots and boundaries. These attention mechanisms enhance feature representation and localization, improving accuracy and robustness. Experimental evaluations show that the method achieves high segmentation performance of dice score 90.01%, surpassing state-of-the-art methods and demonstrating its effectiveness for clinical dental applications.**

***Keywords*—** *CBCT, Squeeze and Excitation attention, Dilated convolution, 3D segmentation, Tooth Centroid, Tooth Skeleton.*

I. Introduction

Three-dimensional (3D) tooth segmentation plays a pivotal role in modern dentistry, enabling precise planning and improved outcomes in orthodontics, dental surgeries, and diagnostics [1]. Unlike traditional 2D imaging, which provides limited perspectives, 3D segmentation allows for the detailed analysis of individual teeth and their surrounding structures. This capability is essential for tasks such as implant placement, orthodontic appliance design, and detecting dental conditions like bone loss or root fractures. By providing a comprehensive view of dental anatomy, 3D segmentation [2] helps clinicians tailor treatments, minimize risks, and enhance overall care quality as shown in Figure 1.

With the rapid advancements in deep learning, segmentation techniques have proven to be highly effective and reliable in medical imaging applications, including dental image analysis. As a result, several studies have explored their application for individual tooth segmentation [3-5]. For instance, Cui et al. [6] introduced ToothNet, a pioneering framework for tooth segmentation in CBCT images. Their approach begins by extracting tooth edges from the input CBCT image and then combines the edge map with the original image using a region proposal network to produce segmentation results.

Similarly, Lee et al. [7] and Gerhardt et al. [8] proposed a two-stage segmentation strategy. In their methods, the initial stage involves localizing individual teeth within the CBCT image using object detection and related techniques. Additionally, Cui et al. [9] developed a fully automatic AI system for simultaneous tooth and alveolar bone segmentation from CBCT images. The method effectively addresses both soft and hard tissue challenges, making it suitable for complex

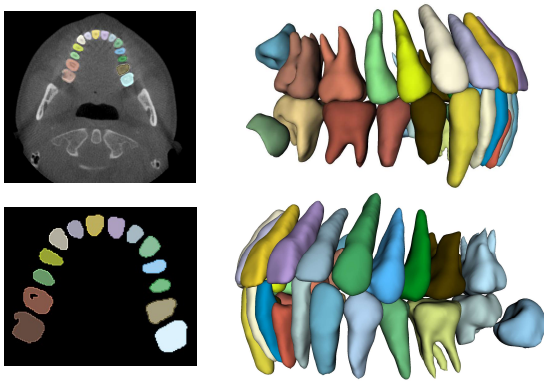

Figure 1: An example output of our teeth segmentation architecture. The upper left is the ground tooth CBCT, bottom left is the predicted segmentation in 2D, top and bottom right are the 3D segmentation of the corresponding image.

dental anatomies. However, its three-stage architecture is highly complex, relying on ROI generation, and occasionally produces inconsistent tooth classifications. More recently, Lv, W., et al. [10] presented a skeleton-guided V-Net for tooth extraction, which is designed to extract the tooth skeleton. However, it was trained on a very small dataset and does not account for metal artifacts, limiting its ability to generalize effectively across different cases.

Previous methods for 3D tooth segmentation often face limitations due to being trained on small datasets, which reduces their generalizability and robustness. Additionally, these approaches struggle with challenging scenarios such as the presence of metal artifacts in CBCT images and cases with closed bite positions, resulting in reduced segmentation accuracy. Furthermore, these methods typically fail to reliably delineate complex dental structures, particularly in areas with intricate root morphologies or low contrast between teeth and surrounding tissues.

To address these shortcomings, we propose a novel learning-based framework for automatic 3D tooth instance segmentation from CBCT images. Our approach incorporates domain-specific adaptations, such as centroid and skeleton-guided segmentation, to enhance precision even in challenging conditions like metal artifacts and closed bite positions. By leveraging shape and data priors, the proposed method effectively isolates teeth from surrounding tissues and distinguishes individual teeth with improved accuracy and reliability.

This study specifically focuses on overcoming these challenges by delivering a robust solution for CBCT-based 3D tooth segmentation, making it more suitable for complex clinical scenarios in modern dentistry.

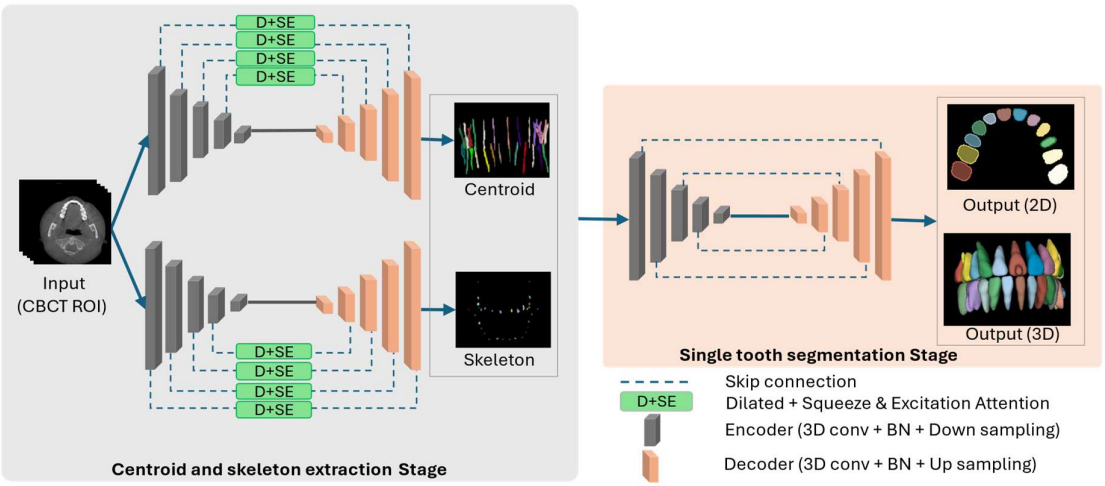

Figure 2: Proposed architecture of our teeth segmentation Network. The left block consists of centroid and skeleton extraction networks. In the skip connection Dilated attention and then Squeeze and Excitation attention (D+SE) is used. The right block does the identification and segmentation of each individual tooth. The result is a 2D output as well as 3D.

II. Proposed Methodology

Figure 2 illustrates the detailed architecture of our proposed method, tailored for high-precision 3D tooth segmentation from CBCT data. The approach consists of two primary stages.

A. Centroid and Skeleton Extraction Stage:

The first stage processes the CBCT to extract crucial structural information. This involves the localization of individual tooth centroids and the generation of a detailed tooth skeleton. The extraction block is built on a V-Net [11] backbone augmented with Dilated attention and Squeeze-and-Excitation attention(D+SE) blocks [12,13], integrated into the skip connections. These attention mechanisms enhance the network's ability to focus on both global and local anatomical features, improving feature representation for accurate centroid and skeleton detection.

1) Centroid Extraction:

This segment of the network predicts the centroid of each tooth by calculating the distance and direction (offset) from very point in the image to the closest tooth center. The result is a map called the centroid offset map, where each point points toward the center of a tooth. This helps locate teeth accurately, which is important for tasks like aligning teeth in orthodontics.

2) Skeleton Extraction:

At the same time, another segment of the network focuses on the shape and structure of the teeth. It creates a binary map that separates teeth from the background and an offset map that shows the distance and direction to the closest point on the tooth skeleton. Using this information, the network generates a detailed skeleton map that highlights the structure of the teeth. This map is further refined to keep only the most important parts, giving a clean and accurate representation of the tooth skeleton.

B. Single Tooth Segmentation Stage:

Once the centroids and skeletons are found, the second stage uses these information to segment each tooth individually. The single tooth segmentation stage uses two-channel inputs comprising patches cropped from the centroid map and skeleton map. Each input patch, sized 96×96×96, encapsulates the anatomical and spatial context of the tooth. These inputs are processed by a V-Net architecture, specifically designed to segment individual teeth with high precision. The network outputs instance segmentation results for each tooth, significantly improving the accuracy of tooth location and providing a detailed analysis of their structural topology. This stage process ensures precise and reliable segmentation of each tooth, even in challenging cases with closely packed or overlapping teeth.

We chose not to use D+SE attention at this stage because we found that it is primarily effective in the initial stages of feature extraction, where it captures spatial and channel relationships comprehensively and efficiently. Applying it at the final stage, however, resulted in overfitting. This issue arises because the deeper architecture of D+SE tends to excessively refine features at later stages, leading to reduced generalizability and a negative impact on the model's overall performance.

C. Detailed Structure of the Dilated Attention and Squeeze-and-Excitation Attention

The Dilated Attention and Squeeze-and-Excitation (D+SE) Block presented in Figure 3, combines dilated convolution-based attention and spatial attention mechanisms, designed to enhance the network's focus on both spatial and contextual features. This block processes 3D volumetric input map $X \in R^{C \times H \times W \times D}$ with tensor dimensions represented as (C,H,W,D) where C represents the number of channels, and (H,W,D) are the spatial dimensions of the input volume.

1) Dilated Attention Mechanism

The dilated attention mechanism employs dilated convolutions to expand the receptive field while maintaining computational efficiency, enabling the capture of both local and global spatial dependencies. In this block, the input map X, with dimensions (C, H, W, D), is passed through a series of 3D dilated convolutions with different dilation rates (1,3,5,7 and 9). These convolutions use a kernel size of 3 and various padding values to preserve the spatial dimensions (H, W, D). The outputs from all convolutions are summed elementwise,

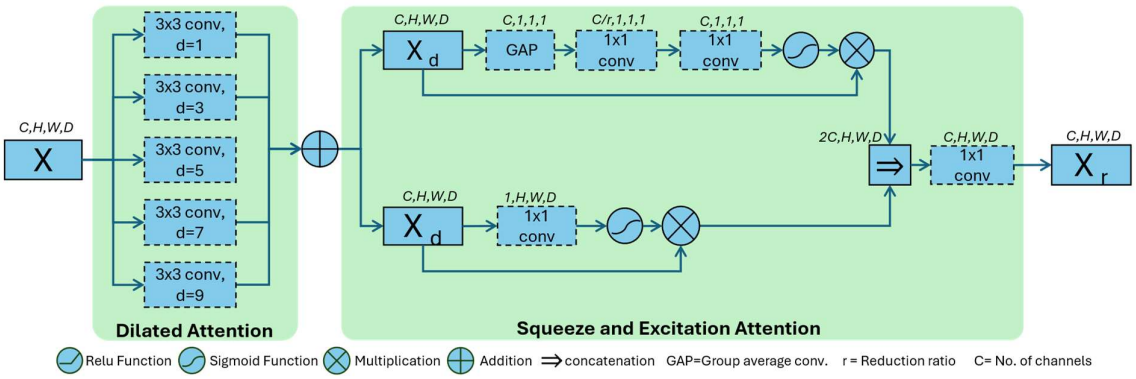

Figure 3: The block diagram of D+SE attention mechanism used in the skip connection of the V-Net architecture. The dilated attention mechanism is shown at the left side, while the Squeeze and Excitation attention mechanism is shown at the right side.

producing a refined map after dilated attention X_d of size (C, H, W, D) same as the input map.

2) Squeeze-and-Excitation (SE) Mechanism

The squeeze-and-excitation (SE) attention block recalibrates channel-wise and spatial-wise features to highlight the most relevant channels. Two types of SE mechanisms are employed within this block: channel attention and spatial attention.

The channel attention operates by first applying an adaptive group average pooling (GAP) operation on the input map X_d, with dimension (C,H,W,D), reducing the spatial dimensions to a map of size (C,1,1,1). This is followed by a 1x1 3D convolution, which acts similarly to fully connected layers, reducing the dimensionality by a factor of r (in our case r=16) and then restoring it back to the original number of channels C. A sigmoid activation function is applied to generate channel-wise attention weights, which are then multiplied elementwise with the original input map, recalibrating the features. The output retains the original dimensions (C,H,W,D), now with enhanced channel-wise attention.

Simultaneously, the spatial mechanism focuses on recalibrating spatial features. A 3D convolution with a kernel size of 1 is applied to the input map, producing a single channel map. After passing through a sigmoid activation function , spatial attention weights are generated, which are then multiplied element-wise with the input map to highlight important spatial regions. The outputs from both the channel and spatial attention mechanisms are concatenated, resulting in a feature map of size (2C,H,W,D). After concatenation, a 1x1 3D convolution is applied to reduce the resulting dimensions back to (C,H,W,D). This final output X_r is the refined output of the input feature map.

D. Loss function

During the training stage, we employ a combined loss function consisting of Binary Cross-Entropy (BCE) Loss and Dice Loss to guide the segmentation process. This approach balances pixel-wise classification and the overlap between the predicted and ground truth segmentation masks. Given X as the input volume batch, G as the ground truth labels, P as the predicted output probabilities after applying the SoftMax function, the BCE loss (L_{BCE}) is presented as (1).

$$L_{BCE}(P,G) = -\frac{1}{C}\sum_{i=1}^{N}\sum_{j=1}^{C} G_{ij}\, log(P_{i,j}) \tag{1}$$

while the Dice loss (L_d) is presented as equation (2):

$$L_d(Y,G) = 1 - \frac{2}{C}\Sigma_{j=1}^{C}\frac{\sum_{i=1}^{N} P_{i,j}G_{i,j}}{\sum_{i=1}^{N} {P^2}_{i,j} + \sum_{i=1}^{N} {G^2}_{i,j}} \tag{2}$$

Here, C is the number of classes, N is the total number of voxels, G_{ij} represents the ground truth probability for the ith voxel belonging to class j, and $P_{i,j}$ denotes the predicted probability for the ith voxel assigned to class j. In the centroid and skeleton component, the combined loss ($L_{c\&s}$) is shown as equation (3):

$$L_{c\&s} = \lambda_c\left(L_{BCE_c}(P,G) + L_{d_c}(P,G)\right) + \lambda_s\left(L_{BCE_s}(P,G) + L_{d_s}(P,G)\right) \tag{3}$$

where λ_c and λ_s represent the contributions of centroid loss and skeleton loss, respectively, both of which are set to 0.5 in our experiments. In the second stage, which involves single tooth segmentation, the same BCE and Dice losses are used to evaluate the discrepancy between the predicted segmentation and the ground truth. The combined loss function for the single tooth segmentation component (L_{st}) is shown as equation (4):

$$L_{st} = \frac{1}{2}L_{BCE_{st}}(P,G) + \frac{1}{2}L_{d_{st}}(P,G) \tag{4}$$

III. Results

A. Implementation Details:

The framework was implemented using the PyTorch library [12], with the Adam optimizer employed to minimize the loss functions and optimize the network parameters through backpropagation. A learning rate of 0.001 and a mini-batch size of 1 were used across all components of the architecture, with a total of 100 epochs. At the end of each epoch, the loss on the validation dataset was computed to monitor network convergence. If the model's performance on the validation dataset showed no improvement over 5 consecutive epochs, training was considered converged and

Table 1: Evaluation metrics

Metrics	Formulas
IoU	$\frac{\sum_{n=0}^{N-1}(t_n^t \; AND \; p_n^p)}{\sum_{n=0}^{N-1}(t_n^t \; OR \; p_n^p)}$
Dice coefficient	$\frac{2\sum_{n=0}^{N-1}(t_n^t \cdot p_n^p)}{\sum_{n=0}^{N-1} t_n^t + \sum_{n=0}^{N-1} p_n^p}$
Precision	$\frac{TP}{TP + FP}$
Recall	$\frac{TP}{TP + FN}$
N: Total number of voxels t_n^t, p_n^t : Ground truth and predicted binary label for voxel n respectively TP, FP,FN: True positive, False positive, False Negative	

was stopped. All models were trained on an NVIDIA GeForce RTX 3090 GPU.

B. Data preprocessing:

The public dataset of 100 CBCT images obtained via official request from [9], has an axial resolution of 400×400 pixels and a depth of 224 to 368 slices. To ensure consistency, the images were resampled to an isotropic resolution of 0.4×0.4×0.4 mm³, removing anisotropy and maintaining uniform spatial resolution. Resampling was done using scaling factors calculated from the original voxel spacing, with trilinear interpolation for labels and nearest-neighbor interpolation for images.

Next, the intensity of the CBCT images is normalized to the range [0, 1]. To generate training data, 35 random patches of size 128×128×128 are cropped around the alveolar bone in each CBCT image. To improve the model's generalization and robustness to variations, we implemented a set of data augmentation techniques tailored specifically for 3D images.

C. Metrics for Evaluation:

We chose Dice score, IoU, precision, and recall evaluating accuracy, overlap, and the balance between predictions, which are essential for 3D segmentation. Their equations and descriptions are in Table 1.

D. Qualitative Comparison

Figure 4 presents a qualitative comparison of tooth identification and segmentation results using a standard CBCT case. The first column illustrates the 2D view of the dental arch, while the subsequent columns display reconstructed 3D tooth models from two perspectives, right and left views. The ground truth segmentation is shown in the first row, followed by results from three competing methods: the skeleton-guided method [10], ToothNet [6], and Automatic AI [9]. Our proposed method is depicted in the final row.

The red rectangles highlight segmentation inaccuracies observed in the alternative methods, particularly in delineating individual teeth and handling overlapping regions. While the skeleton-guided approach struggles with over-segmentation and lacks precision in boundary delineation, ToothNet and Automatic AI exhibit issues such as missed tooth parts and inaccurate separation of adjacent teeth. None of these methods segmented wisdom teeth.

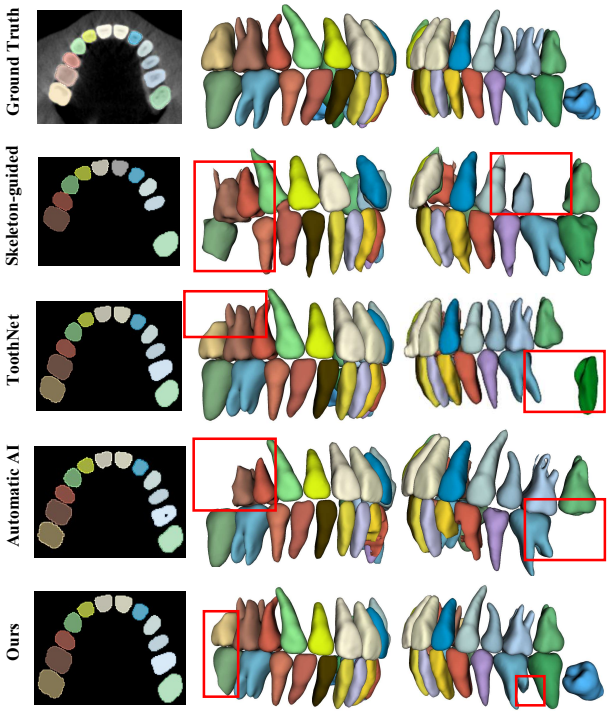

Figure 4: Qualitative comparison of tooth identification and segmentation using a standard CBCT case. The leftmost column displays the 2D view, while the following columns present the reconstructed 3D tooth models from two distinct angles left and right. Red rectangles are used to emphasize segmentation errors in the results.

Table 2: Quantitative comparison.

Methods	Dice %	IoU %	Precision %	Recall %
ToothNet [6]	83.94	73.18	80.70	88.78
Automatic AI [9]	85.10	74.44	90.21	80.90
Skeleton guided [10]	79.91	68.02	77.54	85.23
Ours	**90.35**	**82.61**	**91.43**	**89.78**

Bold represent highest score in each column.

In contrast, our method achieves superior segmentation results, accurately isolating individual teeth and maintaining the integrity of their 3D structures. All the teeth are successfully identified and segmented. The detailed tooth boundaries and minimal segmentation errors demonstrate the robustness of our approach, even in challenging cases involving complex dental anatomies. This qualitative analysis emphasizes the effectiveness of our model compared to state-of-the-art techniques.

E. Quantitative Comparison

Table 2 provides a quantitative comparison of our method with the state-of-the-art methods. As shown in the table, our method outperforms all competing approaches, demonstrating superior accuracy in all metrics. This indicates its effectiveness in addressing the challenges of tooth segmentation, such as distinguishing individual teeth and accurately capturing tooth boundaries. The results clearly highlight the robustness and reliability of our model compared to existing methods.

IV. Conclusion and Future Work

In this study, we proposed a deep-learning framework for 3D tooth instance segmentation from CBCT images, which effectively addresses the challenges posed by complex dental anatomies, including variable tooth sizes, intricate root structures, and metal artifact. Our approach integrates a combination of both dilated attention and squeeze-and-excitation attention mechanism within the skip connections of a V-Net architecture, significantly improving segmentation accuracy. Quantitative and qualitative evaluations demonstrated that our method outperforms state-of-the-art approaches, achieving superior results in segmenting individual teeth and tooth roots. These findings highlight the potential of our model for advancing digital dentistry and improving clinical workflows.

For future work, we aim to enhance the generalizability of our model by training it on a larger and more diverse dataset, including cases with severe pathologies and artifacts. Additionally, we plan to explore the integration of semi-supervised or unsupervised learning techniques to leverage unlabeled CBCT datasets.

V. Acknowledgement

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00217471) and by the National Research Foundation of Korea (NRF) funded by the Korean government under Grant 2022R1I1A3065473.

REFERENCES

[1] H. Wang, J. Minnema, K. J. Batenburg, T. Forouzanfar, F. J. Hu, and G. Wu, “Multiclass CBCT image segmentation for orthodontics with deep learning,” J Dent Res, vol. 100, no. 9, pp. 943–949, 2021.

[2] B. M. Elgarba, S. Van Aelst, A. Swaity, N. Morgan, S. Shujaat, and R. Jacobs, “Deep learning-based segmentation of dental implants on cone-beam computed tomography images: A validation study,” J Dent, vol. 137, p. 104639, 2023.

[3] J. Priya, S. K. S. Raja, and S. U. Kiruthika, “State-of-art technologies, challenges, and emerging trends of computer vision in dental images,” Comput Biol Med, vol. 178, p. 108800, 2024.

[4] S. Jia, G. Wang, Y. Zhao, and X. Wang, “Accuracy of an autonomous dental implant robotic system versus static guide-assisted implant surgery: A retrospective clinical study,” J Prosthet Dent, Jun. 2023, doi: 10.1016/j.prosdent.2023.04.027.

[5] M. Tarce, Y. Zhou, A. Antonelli, and K. Becker, “The Application of Artificial Intelligence for Tooth Segmentation in CBCT Images: A Systematic Review,” Applied Sciences, vol. 14, no. 14, p. 6298, 2024.

[6] Z. Cui, C. Li, and W. Wang, “ToothNet: automatic tooth instance segmentation and identification from cone beam CT images,” in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 6368–6377.

[7] J. Lee, M. Chung, M. Lee, and Y.-G. Shin, “Tooth instance segmentation from cone-beam CT images through point-based detection and Gaussian disentanglement,” Multimed Tools Appl, vol. 81, no. 13, pp. 18327–18342, 2022.

[8] M. do N. Gerhardt et al., “Automated detection and labelling of teeth and small edentulous regions on cone-beam computed tomography using convolutional neural networks,” J Dent, vol. 122, p. 104139, Jul. 2022, doi: 10.1016/j.jdent.2022.104139.

[9] Cui, Zhiming, Yu Fang, Lanzhuju Mei, Bojun Zhang, Bo Yu, Jiameng Liu, Caiwen Jiang et al. "A fully automatic AI system for tooth and alveolar bone segmentation from cone-beam CT images." Nature communications 13, no. 1 (2022): 2096.

[10] Lv, Wenjing, Qunwen Niu, Lanying Wang, Shuang Song, Jiahao Huo, Lin Wang, and Ruoxiu Xiao. "Tooth Segmentation from Cone Beam Computed Tomography using Skeleton-guided V-Net." In Proceedings of the 2024 4th International Conference on Bioinformatics and Intelligent Computing, pp. 297-302. 2024.

[11] F. Milletari, N. Navab, and S.-A. Ahmadi, “V-net: Fully convolutional neural networks for volumetric medical image segmentation,” in 2016 fourth international conference on 3D vision (3DV), Ieee, 2016, pp. 565–571.

[12] Li Y, Zhang X, Chen D. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. InProceedings of the IEEE conference on computer vision and pattern recognition 2018 (pp. 1091-1100).

[13] Hu J, Shen L, Sun G. Squeeze-and-excitation networks. InProceedings of the IEEE conference on computer vision and pattern recognition 2018 (pp. 7132-7141).

[14] A. Paszke et al., “Pytorch: An imperative style, high-performance deep learning library,” Adv Neural Inf Process Syst, vol. 32, 2019.