Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

# Supplementary Material: Frequency-Guided Network for Low-contrast Staining-free Dental Plaque Segmentation

Yiming Jiang[a], Wenfeng Song[b,*], Shuai Li[a,c,**], Yuming Yang[a], Bin Xia[d], Aimin Hao[a], Hong Qin[e,]

[a] *State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing* 100191*, China*
[b] *the School of Computer Science, Beijing Information Science and Technology University, Beijing* 100192*, China*
[c] *Zhongguancun Laboratory, Beijing, China*
[d] *the Department of Pediatric Dentistry, Peking University School and Hospital of Stomatology, Beijing* 100081*, China*
[e] *the Department of Computer Science, Stony Brook University, Stony Brook, New York* 11794-2424*, USA*

ARTICLE INFO

*Keywords:*
Dental plaque segmentation
Oral endoscope images
Convolutional neural network
Multiple tasks framework

ABSTRACT

The current clinical detection of dental plaque relies on medical staining reagents and professional intervention. The deep learning-based automatic segmentation offers an alternative, eliminating the need for medical staining reagents and enabling patients to perform daily plaque detection at home. However, existing methods still suffer from the extremely low-contrast visual features between dental plaque and healthy teeth. To address this issue, we propose a Frequency-Guided Network (FGN) for dental plaque segmentation in extremely low-contrast regions, surpassing human visual capabilities. Considering the characteristic distribution of dental plaque clusters at the junction of teeth and gingiva, our key motivation is to disentangle high-frequency regions and give special attention to plaque in these areas. In addition, we introduce a novel high-to-low frequency multiple tasks segmentation framework that leverages the guidance from high-frequency edges to refine the plaque regions. Our newly designed network employs a robust decoupling and boundary-augmenting paradigm to capture the global clues of the plaque. Through extensive evaluations, our method outperforms existing high-performance segmentation methods. Furthermore, user studies confirm that our approach achieves superior results compared to experienced dentists. https://frequency-guided-network.github.io/

We provide more details for the method and experiments and more quantitative and qualitative results as an extension of the main paper.

## Appendix A. More Method Details

In this section, we give more method details as complementary to the paper's Method parts.

### *Appendix A.1. High-frequency augmentation algorithm*

We conduct a sequence of operations in high-frequency augmentation, including hole-filling, opening, and dilation, to refine the HF-phase output teeth mask, as shown in Fig. A.1.

**The detailed principle of hole-filling.** As teeth surfaces should not contain holes, hole-filling operation [1] helps fill in any holes segmented incorrectly within the teeth region. Hole-filling involves $Marker$, $Mask$ and a structuring element $S_h$. $Marker$ is the starting image of the transformation and will be continuously dilated until it converges. $Mask$ is used to constrain the dilation result, which means $Marker$ cannot dilate more than the $Mask$. The structuring element is used as the kernel of the dilation. The formula about $Marker$, $Mask$ and

[*] Corresponding author:songwenfenga@163.com
[**] Corresponding author:lishuai@buaa.edu.cn

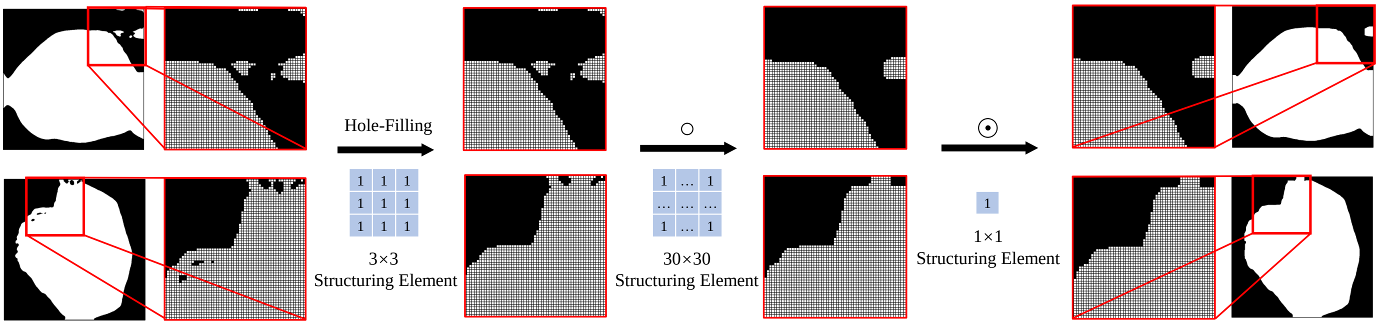


**Fig. A.1. The illustration of the high-frequency augmentation. We conduct the hole-filling operation, opening, and dilation sequentially. Of which, "○" denotes opening while "⊙" denotes dilation. Each small square in the figure represents a pixel.**

structuring element $S_h$ is shown in Eq. A.1.

$$Marker = (Marker \odot S_h) \cap Mask, \tag{A.1}$$

where ⊙ represents dilation, $S_h$ is structuring element, ∩ represents intersection operation.

**The detailed principle of erosion.** Erosion [? ] can be formulated as follows:

$$\begin{aligned} (S_e)_x &= \{y \mid y = a + x, a \in S_e\}, \\ A \ominus S_e &= \{x \mid (S_e)_x \subseteq A\}, \end{aligned} \tag{A.2}$$

where $A$ represents the initial mask, $(S_e)_x$ represents translating all elements of $S_e$ by $x$ units, ⊖ represents the erosion operator, and $S_e$ is the structuring element. The erosion results in translation points make $S_e$ still belong to $A$ after translation. Erosion can also be regarded as calculating the minimum value of the pixel points in the area covered by structuring element $S_e$ and assigning this minimum value to the pixel specified by the reference point.

**The detailed principle of dilation.** Dilation [? ] can be formulated as follows:

$$A \odot S_d = \{x \mid (\hat{S}_d)_x \cap A \subseteq A\}, \tag{A.3}$$

where $A$ represents the initial mask, $S_d$ is the structuring element, $(\hat{S}_d)_x$ means the reversed $(S_d)_x$, ⊙ represents dilation and ∩ represents intersection operation. Dilation can also be regarded as calculating the maximum value of the pixel points in the area covered by $S_d$ and assigning this maximum value to the pixel specified by the reference point.

**The detailed principle of opening operation.** We perform an opening operation [? ] to eliminate noise from the teeth region of the image. As the teeth area is typically a connected and sizable region, small noise could introduce redundant boundaries and adversely affect the LF phase. The opening operation involves erosion followed by dilation, where erosion can be regarded as assigning the minimum pixel value in the area covered by a structuring element to the specified reference point. We previously explained dilation in the hole-filling method. We perform the opening operation with a 30 × 30 structuring element to remove noise without distorting the main body of teeth. The opening operation can be formulated as follows:

$$A \circ S_o = (A \ominus S_o) \odot S_o, \tag{A.4}$$

where $A$ represents the initial mask, ∘ represents the opening operator, and $S_o$ is the structuring element.

### *Appendix A.2. Frequency-guided decoupling*

The proposed method for disentangling the high-frequency part and providing independent supervision is illustrated in Fig.A.2.

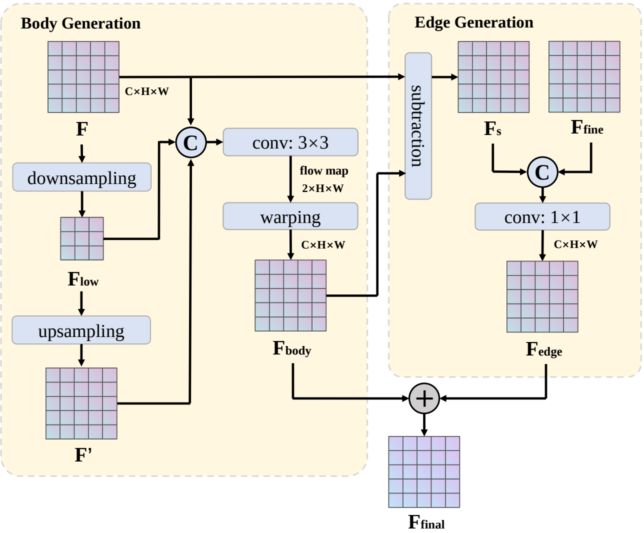


**Fig. A.2. The illustration of frequency-guided decoupling. Of which, "Ⓒ" denotes concatenation and "⊕" denotes element-wise addition.**

### *Appendix A.3. Attention mechanism*

**The channel attention module.** The channel attention maps [2] capture the inter-channel relationship of the features. The input feature map is subjected to global max pooling and global average pooling based on width and height, generating two feature maps with the size of $1 \times 1$, and the number of channels consists of the original feature map. Then, the two feature maps are sent to a shared two-layer neural network composed of multi-layer perception (MLP) with one hidden layer. The MLP contains two layers: $W_0 \in \mathbb{R}^{C/r \times C}$ and $W_1 \in \mathbb{R}^{C \times C/r}$, $r$ is the reduction ratio, which is set to 16 in our network. The *ReLU* activation function is followed by $W_0$. After that, the MLP output features are subjected to an element-wise addition, and then a *sigmoid* function is performed to generate the final channel attention map.

**The spatial attention module.** The spatial attention module [2] generates the spatial attention map by using the interspatial relationship of the features. First, the input feature map

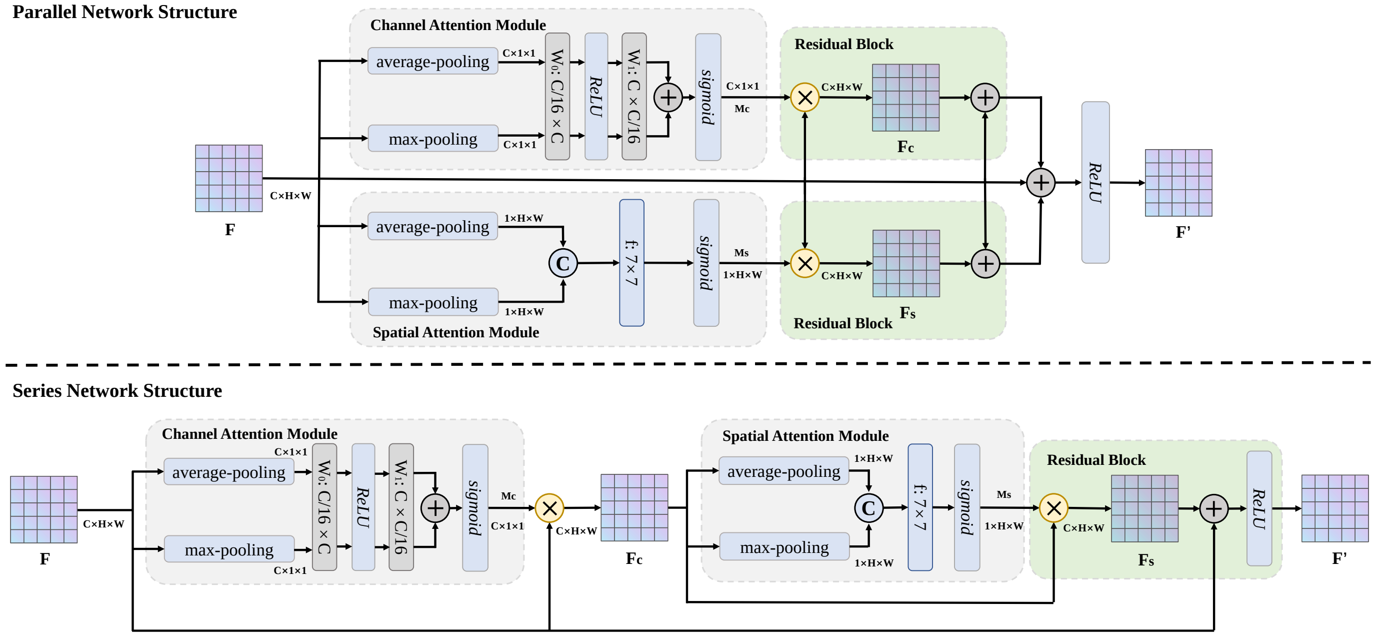


**Fig. A.3. The comparison of the parallel and series network structure of the attention mechanism in our FGN. We integrate channel and spatial attention modules into our method by adding them as residual blocks in parallel. Of which, the "⊕" symbol denotes element-wise addition, while the "⊗" symbol denotes the Hadamard product. The gray box indicates the convolution layer, while the blue box with a blue border signifies a convolution operation with a filter.**

is applied global max pooling and global average pooling based on the channel, yielding two feature maps in $\mathbb{R}^{1\times H\times W}$. Then they are concatenated by channels, and the channel will be reduced to 1 after a convolution operation with a filter whose size is $7 \times 7$. At last, the module will generate the spatial attention feature through the *sigmoid* function.

**Attention modules combination.** Two possible methods of combining the channel and spatial attention modules are in parallel or in series. To illustrate their differences, we compare the parallel and series network structures in Fig. A.3. After performing preliminary experiments, we include the channel and spatial attention modules as a residual block in parallel.

## Appendix B. More Experiment and Evaluation Details

This section gives more details as complementary to the paper's Experiment and Evaluation parts.

### *Appendix B.1. Parameter analysis*

We analyze the critical parameters of high-frequency augmentation and the attention mechanism positions.

**The high-frequency augmentation parameters.** The high-frequency augmentation parameters affect the LF phase's input, influencing our methods' performance. The two main parameters are $S_d$ and $S_o$. To investigate the impact of these parameters, we perform high-frequency augmentation with three different sets of parameter values on the teeth mask. This results in three datasets, namely ENHAN-1, ENHAN-2, and ENHAN-3. Then, we evaluate our methods on each dataset as the input of the LF phase to assess the mIoU and mAcc. The result are shown in Table B.1, We observe that the ENHAN-1 dataset performs slightly better than the other two. As a result, we empirically select the parameter values from ENHAN-1 for our methods.

**Table B.1. Performance comparison of three structuring element sizes.**

| Datasets | $S_o$ | $S_d$ | Methods | mIoU[%] | mAcc[%] |
|---|---|---|---|---|---|
| ENHAN-1 | 30×30 | 1×1 | Ours-H<br>Ours | *73.30*<br>**74.02** | *83.75*<br>**84.04** |
| ENHAN-2 | 30×30 | 3×3 | Ours-H<br>Ours | 73.22<br>73.77 | 83.35<br>83.78 |
| ENHAN-3 | 15×15 | 3×3 | Ours-H<br>Ours | 73.19<br>73.89 | 83.31<br>83.81 |

**Table B.2.**
**Performance comparison of attention modules in four positions.**

| Attention Modules Positions | Datasets | mIoU[%] | mAcc[%] |
|---|---|---|---|
| No attention module | ENHAN-1<br>ENHAN-3 | 73.48<br>73.30 | 83.33<br>83.50 |
| ASPP | ENHAN-1<br>ENHAN-3 | 73.38<br>73.45 | 83.50<br>83.44 |
| Edge part | ENHAN-1<br>ENHAN-3 | 73.87<br>73.94 | 84.02<br>83.84 |
| The last feature map | ENHAN-1<br>ENHAN-3 | **74.02**<br>73.89 | **84.04**<br>83.81 |

**The attention modules positions.** Furthermore, we investigate the impact of adding the attention modules at different positions in our method. We add the channel and spatial attention modules after certain positions in parallel. Specifically, we consider four different positions for adding the attention modules: no attention module, after the ASPP module, after the edge part of the decoupling module, and after the last feature map before

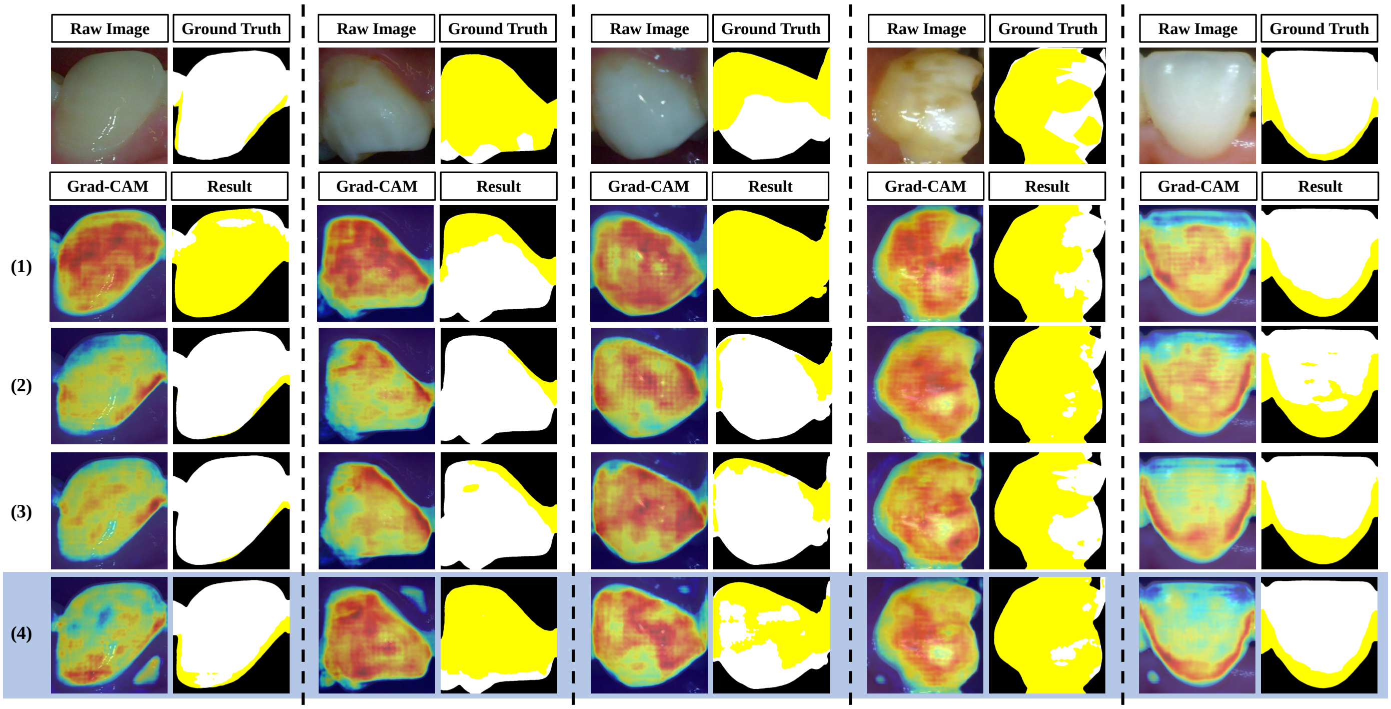


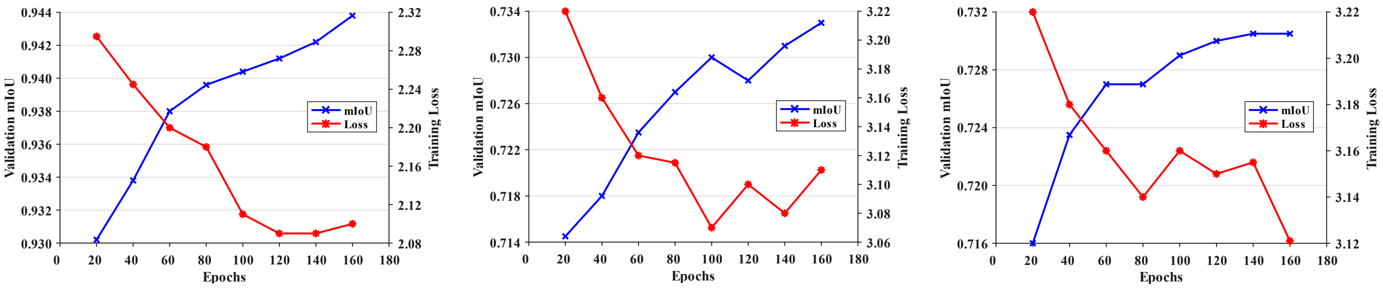
**Fig. B.4. Qualitative comparisons with four attention modules positions. In the top row, we demonstrate the raw image and ground truth for reference. We conduct experiments with four different positions for adding the attention modules. In the second to fifth rows, we show the results and Gradient-weighted Class Activation Mappings [3] of the final layer. Although some CAM diagrams exhibit the checkerboard pattern due to grad-CAM's limitations, it does not affect the maps' interpretation. The four positions for the attention modules are as follows: (1) no attention module, (2) attention modules added after the ASPP module, (3) attention modules added after the edge part of the decoupling module, and (4) attention modules added after the last feature map and before the final upsampling.**

the final upsampling. The effect of the attention modules varies for the two variants of our method. Table B.2 shows the results obtained for the four different positions on datasets ENHAN-1 and ENHAN-3. The results are similar for each position on both datasets, but the gaps between different positions are significant. Adding the attention modules after the last feature map results in the best performance, increasing 0.54% in mIoU and 0.69% in mAcc. Fig. B.4 provides visualization results of the final layer's Gradient-weighted Class Activation Mappings (CAM). Notably, adding the attention modules after the last feature map allows the network to focus more on the areas of the teeth not covered by plaque. Although some CAM diagrams exhibit the checkerboard pattern due to grad-CAM's limitations, it does not affect the maps' interpretation.

### *Appendix B.2. Visual analysis and visualization results*

**Visualization results of the training process metrics.** In every training phase, the training loss drops, and the validation mIoU rises with increased epochs. The training loss and validation mIoU of the HF phase are shown in Fig. B.5(a). The training loss and validation mIoU of the LF phase of our methods are shown in Fig. B.5(b) and Fig. B.5(c). In three figures, the blue line shows mIoU, and the red line shows training loss. The left horizontal axis represents validation mIoU, the right horizontal axis represents training loss, and the vertical axis denotes epochs. The loss curve is smoothed using a smoothing rate of 0.99, while the mIoU curve is smoothed with a smoothing rate of 0.9. From these figures, it can be observed that the training loss drops, and the validation mIoU rises with the increase in epochs from 20 to 160. Although the curve of training loss shows large fluctuations after 100 epochs, it is important to note that the loss curve actually drops significantly more from 0 to 20 epochs when compared to the fluctuations seen after 100 epochs. In summary, the curves in all three figures start to fluctuate after 100 epochs, indicating that the model has been trained to its optimal performance.

**Visualization results of the body and edge binary map.** The visualization of the body and edge binary map generated by the HF and LF phases are shown in Fig. B.6. In the HF phase, our methods focus on segmenting the teeth and non-teeth regions, resulting in a binary edge map that accurately identifies the edges of teeth. Our approaches can precisely locate the edge of teeth when some of the boundaries between teeth and gingiva are blurry. In the LF phase, we segment the plaque and non-plaque regions, leading to a binary edge map highlighting the edges of both plaque and teeth. The binary body map is complementary to the corresponding edge binary map, providing additional information on the location of plaque and teeth within the image.



(a) Loss and mIoU of the HF phase.

(b) Loss and mIoU of the LF phase of Ours.

(c) Loss and mIoU of the LF phase of Ours-H.

**Fig. B.5. Training loss and validation mIoU of the HF and LF phases in training. (a) Shows loss and mIoU of the HF phase, (b) shows loss and mIoU of the LF phase of Ours, and (c) shows loss and mIoU of the LF phase of Ours-H.**

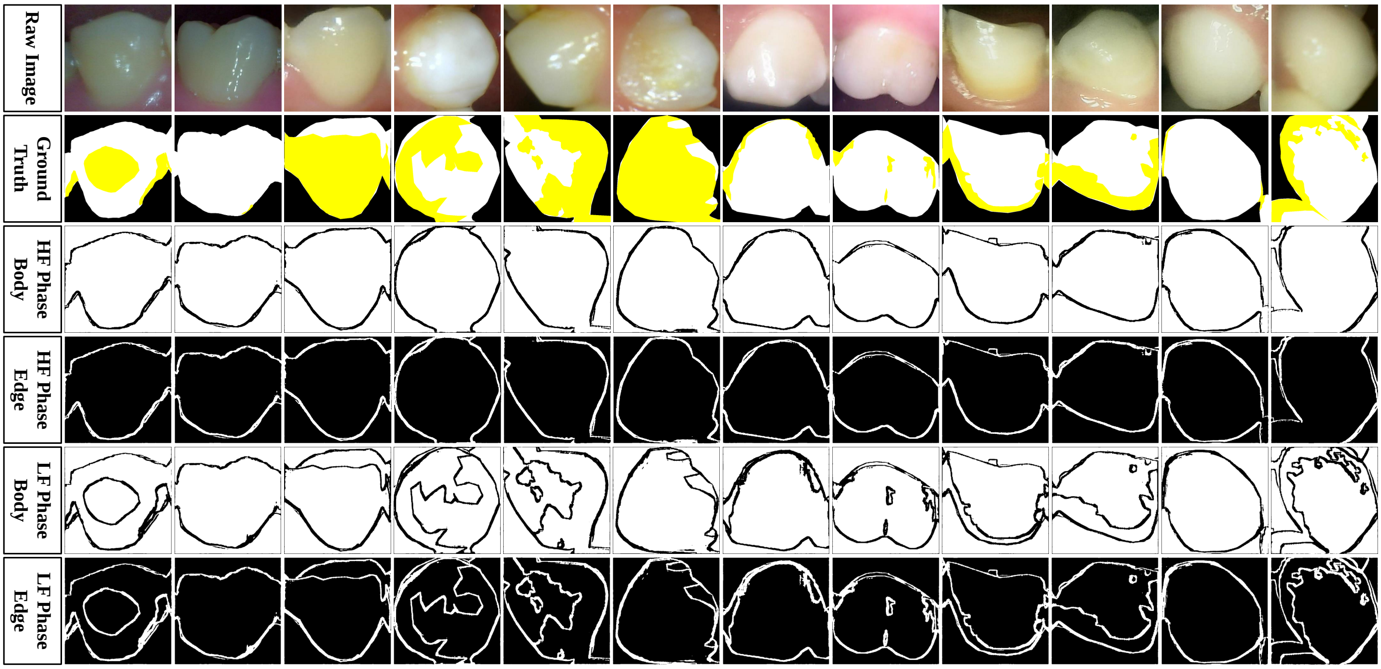


**Fig. B.6. Visualization segmentation results in HF and LF phases for the edge and body. The binary edge maps correspond with the boundary of teeth and plaque precisely. The binary body and edge maps are complementary to each other. The third and fourth rows of the visualization show the binary body map and the binary edge map in the HF phase, while the fifth and sixth rows show the binary body map and the binary edge map in the LF phase.**

## References

[1] R. C. Gonzales and P. Wintz, *Digital image processing*. Addison-Wesley Longman Publishing Co., Inc., 1987.

[2] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “Cbam: Convolutional block attention module,” in *ECCV*, pp. 3–19, 2018.

[3] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” in *ICCV*, pp. 618–626, 2017.