Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/mbe

MBE, 20(2): 2039–2060.
DOI: 10.3934/mbe.2023094
Received: 28 September 2022
Revised: 31 October 2022
Accepted: 03 November 2022
Published: 11 November 2022

Research article

A weakly supervised learning-based segmentation network for dental diseases

Yue Li*, **Hongmei Jin and Zhanli Li**

College of Computer Science and Technology, Xi'an University of Science and Technology, Xi'an 710000, China

* **Correspondence:** Email: yueligzqy0824@163.com; Tel: +8615349218145.

Abstract: With the development of deep learning, medical image segmentation has become a promising technique for computer-aided medical diagnosis. However, the supervised training of the algorithm relies on a large amount of labeled data, and the private dataset bias generally exists in previous research, which seriously affects the algorithm's performance. In order to alleviate this problem and improve the robustness and generalization of the model, this paper proposes an end-to-end weakly supervised semantic segmentation network to learn and infer mappings. Firstly, an attention compensation mechanism (ACM) aggregating the class activation map (CAM) is designed to learn complementarily. Then the conditional random field (CRF) is introduced to prune the foreground and background regions. Finally, the obtained high-confidence regions are used as pseudo labels for the segmentation branch to train and optimize using a joint loss function. Our model achieves a Mean Intersection over Union (MIoU) score of 62.84% in the segmentation task, which is an effective improvement of 11.18% compared to the previous network for segmenting dental diseases. Moreover, we further verify that our model has higher robustness to dataset bias by improved localization mechanism (CAM). The research shows that our proposed approach improves the accuracy and robustness of dental disease identification.

Keywords: weakly supervised semantic segmentation; attention compensation mechanism; conditional random field; joint loss function; class activation map

1. Introduction

Regular screening and clinical diagnosis of dental diseases are mainly evaluated by Experienced dentists through manual probing combined with imaging-assisted analysis. Diagnosis becomes more difficult when more complex situations (such as changes in gingival biotypes) are involved. Optical coherence tomography (OCT) technology is a standard clinical imaging protocol for observing the microstructure of tooth tissue, which is used to identify caries and gingival diseases under complex tissues for more effective prevention procedures [1–4]. However, dental diseases have no obvious symptoms in the early period, which makes it difficult for manual assessment to make accurate diagnoses timely, and is easily influenced by subjective factors such as the doctor's experience. Using computer technology to assist dentists in early diagnosis can help reduce the workload, improve their work efficiency and quality, and ultimately provide better patient services, which is of great clinical significance.

Deep learning, as a complete end-to-end model, can automatically learn nonlinear high-dimensional features of complex data to make the model stronger generalization ability. It has been widely used in computer-aided medical detection and diagnosis [5]. Rana et al. [6] introduce convolutional neural networks (CNN) combined with machine learning classifiers to segment normal and inflamed gingiva from intra-oral images. Dental professionals can get accurate information for the early diagnosis of periodontal disease through this segmentation. Xu et al. [7] focus on dental image segmentation and labeling the dental images using a two-level hierarchical CNN approach. First approach targets teeth-gingiva labeling, and the other targets inter-teeth labeling. This helps to segment the teeth and accurately label all the teeth. Sivagami et al. [8] use the UNet architecture for segmentation of dental x-ray images and obtain 97% accuracy and 94% Dice score. Subsequently, Li et al. [9] propose a few-shot learning method for the intelligent dental plaque segmentation directly using oral endoscope images, which is to conduct few-shot learning at the superpixel level and integrate the superpixels global and local features towards better segmentation results. Kaya and Akar [10] proposed a more sophisticated model distinguishing background, tooth tissue type, and decayed tooth. However, the accuracy of this classification model is poor since decayed tooth tissue is challenging to detect in the early stages.

The deep learning models of the above dental disease segmentation tasks are supervised on private datasets because the oral field has no public dataset. Therefore, it is difficult to establish an objective comparison scale for different studies, and the dataset used for training is too small, which directly limits the generalization ability of the models. The unsupervised deep learning method does not require labeled samples. Self-encoding can also learn an equality function to make the visible layer data and the encoded and decoded data as equal as possible, but its robustness is still poor. In particular, it is less effective in private datasets when the probability distributions of the test and training samples differ significantly [11]. Weakly supervised learning, as a compromise, trains models using image-level labels. It can accurately compute the threshold scores that can trigger the category of pixels and performs better on segmentation tasks [12]. Therefore, This paper proposes an end-to-end semantic segmentation network based on weakly supervised learning to identify dental diseases.

In this paper, we use image-level labels to generate high-quality pseudo labels through the attention mining branch, which largely avoids the dependence on manual labeling. However, in weakly supervised learning using only image-level labels, as shown in Figure 1, the network, when identifying a decayed tooth on the medial side of the posterior alveolar, focuses on the occlusal surface (b) that is highly correlated with it, not on the object itself. In this case, due to the imbalance of the training set, the network learns a priori knowledge that decayed teeth mainly occur on the occlusal surface of the pit and fissure and has no motivation to focus on the foreground (medial posterior alveolar (c)) during training. To address this problem, we propose an ACM based on CAM for self-supervision. This mechanism forces attention to focus on the whole region rather than the most probable occlusal surface by mask erasing to ensure that all regions of interest are included in the attention of the network. For the included non-target region, some mislabeled pixels are removed by CRF constraint checking network attention to produce a reliable network attention map (c), which generates high-quality pseudo labels after a thresholding operation. With limited pixels of pseudo labels as supervision, We introduce a regularized dense energy loss (EN-Loss) combined with pixel cross entropy loss (CE-Loss) to jointly optimize the training process, which can suppress the wrong pixels introduced by cross entropy classification loss and improve the segmentation performance of the network.

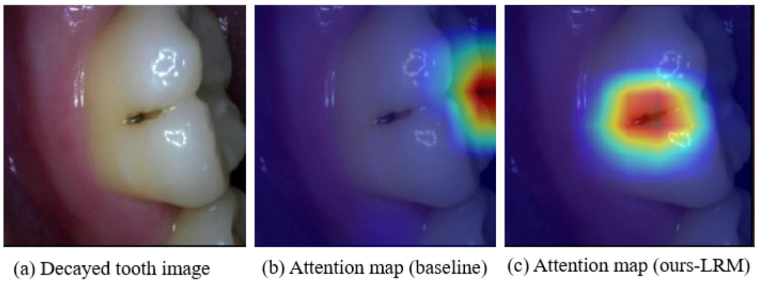


Figure 1. Qualitative localization results of the biased dataset are generated by the CAM [24] technique. The heat map represents the CAM, which locates the attention of the network by highlighting regions. These highlighted regions means higher class activation scores.

Summarily, our primary contributions are as follows: 1) Our use of CAM-Mask erasing allows the network to apply more attention to feature extraction and semantic representation, improving the network's ability to capture more comprehensive features. Thus, the robustness and generalization performance of the model will be enhanced. 2) We use the ACM and CRF post-processing to effectively alleviate the under-activation of foreground regions and over-activation of background regions in weakly supervised semantic segmentation, bridging the gap between the pseudo label and the ground truth. 3) The EN-Loss function we designed considers the utilization of both labeled and unlabeled regions, avoiding the error introduced by CE-Loss and improving the accuracy of semantic segmentation.

The remainder of the paper is organized as follows. Section 2 describes related work, including weakly supervised semantic segmentation and CAM mapping. Section 3 presents a detailed description of the proposed method, and the experimental results and analysis are provided in Section 4. Finally, the conclusions of this process are discussed in Section 5.

2. Related works

2.1. Weakly supervised semantic segmentation

Semantic segmentation is a fundamental computer vision task aiming to predict the pixel-level classification results of images [13]. The fully convolutional network for semantic segmentation was proposed in 2015 [14]. Fully supervised semantic segmentation requires dense pixel-level annotation. Also, manual labeling of datasets is a very time-consuming and labor-intensive task, and the reliability of data labels is closely related to the business capabilities of labeling experts. However, it is indispensable in many computer vision applications (e.g., autonomous driving [15], intelligent medical [16]), which makes weakly supervised learning develop into a new research direction. Weakly supervised semantic segmentation has attracted much attention due to the need for less manual intervention. The so-called weak supervision is to replace pixel-level truth labeling with more easily available truth labeling. Common weakly supervised labels include bounding box, scribble, point, and image-level class labels. Among them, the weakly supervised segmentation based on bounding boxes labeling [17] gives the location and label of the object with a rectangular box (2D) or cuboid (3D), which can effectively distinguish the labeling modes of different instances, including both semantic information and instance information. However, applying it to semantic segmentation scenarios means that some foreground information will be injected, so the effect is not ideal. Therefore, this labeling method is widely used in instance and panoramic segmentation. Scribble labeling [18] does not require dense labels, and just a few lines are drawn to distinguish different regions as supervision information. But the difficulty is that the supervised training region is very small, and only these regions may not provide enough information. The point labeling [19] uses poles as a guide to boxing the object's approximate location. Yet, the poles are difficult to select because they need to fit the object boundaries, especially for some oddly shaped objects. Weakly supervised segmentation based on image-level labeling [20] mainly uses image classification labels to learn the commonality between high-level semantic information of the same class. It is ideal for semantic segmentation and takes the least time and effort, but it cannot distinguish instances. This paper adopts image-level class labels for weakly supervised semantic segmentation.

Image-level weakly supervised semantic segmentation only needs to provide image-level annotation. Most solutions generate initial object seeds or regions based on CAMs, which are transformed to generate pseudo labels to train the semantic segmentation model. Wei et al. [21] propose to iteratively erase and compute discriminative regions of classification networks so that more seed regions can be mined. These regions are then combined with saliency maps to generate pseudo-pixel-level labels. It is also shown that dilated convolution can increase the receptive field and improve the performance of weakly supervised segmentation networks. Subsequently, regional and pixel networks are trained to predict from the image level to the regional level gradually, and from the regional level to the pixel level [22]. In addition, the method also uses saliency maps as additional supervision. Specifically, a traditional seed growth algorithm was implemented to expand the seed region iteratively [23]. However, all the above methods use various methods to generate high-quality pseudo masks, meaning they need at least one or two different networks before training for semantic segmentation prediction. In this research, we design an end-to-end branch network by sharing the backbone parallel to simplify the process.

2.2. *CAM*

Generating the CAM [24] using the CNN classification model plays a vital role in computer vision. Grad-CAM [25] extends CAM and can be applied to any CNN-based classification model. Grad-CAM technology has been widely applied in other weakly supervised vision tasks, such as localization [26], detection [27], and segmentation [28]. Grad-CAM uses the gradient mean of channels obtained in backpropagation as the channel weight, and the generated CAM has a similar effect.

CAM is usually used as a visualization technique in image classification tasks but is rarely used to provide feedback information into the network during training. The top-down attention method based on CAM is proposed to locate the most discriminative object regions in the image. Then, the mined areas are erased from the original image, and the erased image is retrained to another classification network to locate other object regions. This process is repeated to train until the network does not converge well on the erased image. Finally, the erased areas are integrated as the mined object regions [21]. Wang et al. [28] introduced a pixel correlation module (PCM), which exploits context appearance information and refines the prediction of current pixels by its similar neighbors, leading to further improvement in CAMs consistency. These methods depend on complex network structures and have high computational costs. In contrast, in our research, inspired by the idea of CNN erasing to see more, CAM based on gradient propagation (Grad-CAM [25]) is used to generate attention masks. Then the original image is erased and sent to the attention mining branch for loss calculation, equivalent to complementary supervision with the attention mechanism used in reverse, which has the advantage of focusing on the whole rather than the most discriminative regions.

3. Proposed method

3.1. Overview

The overall framework is shown in Figure 2. Our method is named lesion region mining (LRM), including the attention mining branch and segmentation branch. The former uses image-level annotation to generate pseudo-pixel-level masks, and the latter generates semantic segmentation results. Specifically, the attention map generated by the CAM based on gradient propagation (Grad-CAM [25]) is first used as the original seed region. Then the foreground pixel of the seed region is expanded by the ACM, and the CRF operation [29] constrains the background pixels of the seed region to generate the pseudo-pixel-level masks. Finally, these pseudo labels are fed into the segmentation branch for training. The joint loss function and the standard backpropagation algorithm jointly optimize the segmentation model in an end-to-end way.

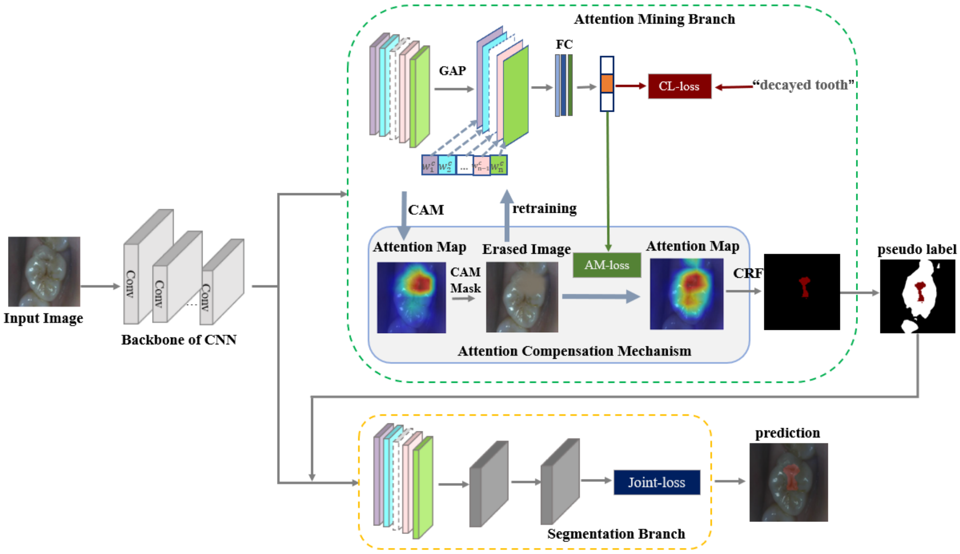


Figure 2. The network framework for segmentation of dental diseases based on weakly supervised learning.

3.2. *Attention mining branch: generating pseudo labels*

The attention mining branch aims to generate reliable and high-confidence pseudo-pixel-level labels using image-level annotations. The original CAM can highlight the most discriminative regions of the object but still contains some non-target regions, that is, mislabeled pixels. Therefore, after obtaining the original CAM regions, we expand foreground pixels and suppress background pixels through our ACM combined with dense CRF post-processing. Finally, high-quality pseudo-pixel-level labels are generated for training and inferring the segmentation branch.

3.2.1. Original CAM

In multi-label scenarios (such as target = [dental calculus, decayed tooth, gingivitis]), it is easy to have an extreme imbalance in the number of positive and negative samples for a single category. In this case, the effect of BCELoss for multi-label classification will be poor. Therefore, in the attention mining branch, our CL-Loss uses the sigmoid function combined with the BCELoss function to perform multi-label classification. It supports the mixed segmentation task with a sample containing multiple labels. The specific expression is shown in Eq (3.1).

$$L_{CL} = \frac{1}{1+e^{-x}} + L_{BCE} \tag{3.1}$$

$$L_{BCE} = -\frac{1}{C}(\sum_{i}(y_i * lnx_i + (1-y_i) * ln(1-x_i))) \tag{3.2}$$

where x_i is the label predicted by the model, its shape is (N,C), N represents the batch size, and C is the number of classifications. y_i is the ground truth label.

We update the network by backpropagation, calculate the contribution value of each pixel space in the original region, and generate the original attention map combined with the CAM technology.

3.2.2. ACM

The ACM uses the information predicted by the network to generate CAM-Mask for erasing. The erased images are retrained by loss calculation. The leading mind of this mechanism is that the original image obtains a higher classification score, and the erasing action tries to mask the valuable features. Hence, the erased image achieves the lower classification score as much as possible. Finally, the model is guided by the ACM, and all foreground regions that contribute to the classification decision will be included in the attention of the network. Overlabeled foreground pixels will be suppressed by CRF validation, making the attention map more accurate.

CAM-Mask: We used CAM based on gradient propagation to simplify the generation of the attention map. In the attention mining branch, for a given image I, let $f_{l.k}$ represents the activation function of unit K in the l-th layer. For each class c from the ground truth label, we calculate the gradient of the score y^c corresponding to class c by deriving the activation map of $f_{l.k}$. These gradients will be fed back through a global average pooling layer to obtain important neuron weights as defined in Eq (3.3).

$$w^c_{l,k} = GAP\left(\frac{\partial y^c}{\partial f_{l.k}}\right) \tag{3.3}$$

where *GAP* represents the global average pooling layer operation.

We do not directly update the network parameters in the training process after $w^c_{l,k}$ is obtained by backpropagation. Since $w^c_{l,k}$ represents the importance of the activation map for the prediction of class c, we use the weight matrix w^c as the convolution kernel and apply 2D convolution on the activation map matrix f_l to integrate all the activation maps. Then a ReLU operation follows to obtain the attention map A^c with Eq (3.4). The attention map is trained online, and the constraint of A^c will affect the network's learning.

$$A^c = ReLU\,(Conv(f_l, w^c)) \tag{3.4}$$

where l is the representation from the highest convolution layer, which has a good balance between high-level semantics and detailed spatial information [30]. Considering the attention map has the same size as the convolution feature map, we directly use the trainable attention map A^c to binarize and generate an attention mask (CAM-Mask), which is used as an image mask to erase valuable features of the original image. The erasing method is shown in Eq (3.5). I^c is the erased image using the CAM-Mask.

$$I^c = I - (T(A^c) \odot I) \tag{3.5}$$

where c is the category, I is the original image, and $T(A^c)$ is the threshold binarization operation. $\odot$ represents cell multiplication, multiplied pixel-wise. To make it derivable, we use the sigmoid function as an approximation, as shown in Eq (3.6).

$$T(A^c) = \frac{1}{1 + e^{-\omega A^c + \sigma}} \tag{3.6}$$

where σ is a threshold matrix whose elements are all equal to σ. ω is a scale parameter, ensuring that $T(A^c)_{i,j}$ are approximately equal to 1 when larger than σ, or to 0 otherwise.

Attention mining loss: The CAM-Mask acts on the original image to generate I^c, fed into the baseline network for training to obtain class prediction scores. Since our goal is to guide the network to focus on all the target regions of interest, and I^c is the erased image without the most discriminative regions. We should ensure that I^c contains as few features as possible belonging to the target class. In the ideal case, other regions should not include a single pixel triggering the class c except the high-responding region on the attention map. From the perspective of the loss function, an attempt should be made to minimize the prediction score of I^c for class c. Therefore, we design a loss function named attention mining loss, as shown in Eq (3.7).

$$L_{AM} = \frac{1}{N}(\sum_c y^c I^c) \tag{3.7}$$

where $y^c I^c$ is the prediction score of I^c for class c. N is the number of class labels for image I.

Through AM-Loss, we can perform online complementary learning on the erased image I^c. If the I^c image still contains some pixels triggering class c objects, the prediction information can be compensated by minimizing the attention mining loss function (AM-Loss).

3.2.3. CRF

The original CAM can highlight the most distinctive regions of the object, but they still contain some non-target pixels mislabeled. Therefore, after obtaining the original CAM region, CRF is needed for some post-processing. The first step is to normalize the CAM score to obtain the classification probability of each foreground pixel in image I, as shown in Eq (3.8).

$$P^c_{fg} = \frac{I^c_{CAM}}{\max(I^c_{CAM})} \tag{3.8}$$

The background score is calculated using a similar way as in Eq (3.9).

$$P^c_{bg} = (1 - \max_{1<x<100}(P^c_{fg}(i)))^\gamma, \gamma > 1 \tag{3.9}$$

where i is the pixel position index, γ is the decay rate factor that suppresses background labels. The overall probability map, namely $P_{fg,bg}$, is obtained by concatenating foreground and background probabilities P^c_{fg} and P^c_{bg}.

After that, we introduce the dense CRF as post-processing to remove some mislabeled pixels, and the CRF pixel label map is:

$$I_{crf} = CRF(I, [P_{fg}, P_{bg}]) \tag{3.10}$$

The learned reliable CAM label is:

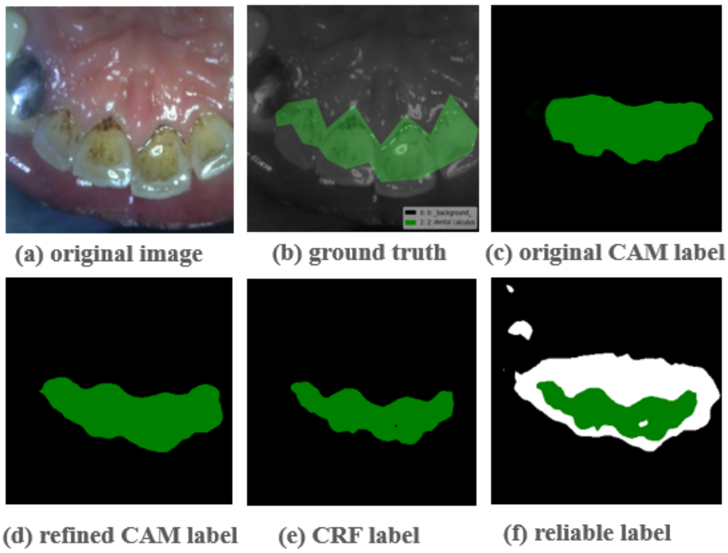
$$I_{cam}(i) = \begin{cases} \underset{c\in C}{max}(P^c_{fg,bg}(i)), & \underset{c\in C}{max}(P^c_{fg,bg}(i)) > \alpha \\ 255, & other \end{cases} \tag{3.11}$$

where C={c_0,c_1,...,c_N} contains all classes and the background (c_0). 255 indicates the class label is not decided yet. $\underset{c\in C}{max}(P^c_{fg,bg}(i)) > \alpha$ means the selection of high confidence region. The final pixel label input to the segmentation branch is:

$$I_f(i) = \begin{cases} I_{cam}(i), & I_{cam}(i) = I_{crf}(i) \\ 255, & other \end{cases} \tag{3.12}$$

where $I_{cam}(i) = I_{crf}(i)$ considers the CRF constraints. Taking this strategy, highly reliable regions, as well as ground truth labels, can be obtained. The regions detonated as 255 are regarded as unreliable regions.

Figure 3 shows an example of our approach. It is observed that the original CAM label (Figure 3(c)) only contains foreground pixels of the most discriminative regions, while the improved CAM label (Figure 3(d)) contains overall foreground pixels but introduces many background pixels as foreground. The CRF label (Figure 3(e)) can get an accurate boundary, but at the same time, some foreground pixels are regarded as background. In other words, the CAM label can provide reliable background pixels, and the CRF label can provide reliable foreground pixels. Combining the CAM label and CRF label map using our method (Figure 3(f)) removes some wrong pixel-level labels but still retains reliable regions.



(a) original image (b) ground truth (c) original CAM label

(d) refined CAM label (e) CRF label (f) reliable label

Figure 3. An example of generating reliable pixel labels.

3.3. Semantic segmentation branch : Prediction

After obtaining reliable pseudo-pixel-level labels, input them into the semantic segmentation branch for training. Our segmentation network is optimized by the joint loss function, including CE-Loss and EN-Loss. CE-Loss focuses on labeled regions, while EN-Loss considers both labeled and unlabeled regions. The joint loss is expressed as :

$$L_{Joint} = L_{ce} + L_{en} \tag{3.13}$$

where L_{ce} and L_{en} represent the CE-Loss and EN-Loss, respectively. The CE-Loss is expressed as :

$$L_{ce} = -\sum_{c\in C, i\in\Phi} B_c(i) log(P^c(i)) \tag{3.14}$$

where $B_c(i)$ is the binary indicator, which equals to 1 if the label of pixel i is c and otherwise 0; Φ denotes the labeled region, $\Phi = \{i|I(i) \neq 255\}$; P^c means the output probability of the training network.

Since CE-Loss is designed for supervised learning, in our work, all pixel labels are not 100% reliable, which means using CE-Loss may introduce some errors. Therefore, the soft filter S(i) is designed for pixel i to alleviate the error introduced by CE-Loss.

$$S(i) = \begin{cases} 1 - max_{c\in C}(P^c(i)), & i \in \Phi \\ 1, & other \end{cases} \tag{3.15}$$

We first define the energy formulation between pixel i and j based on [31]:

$$E(i, j) = \sum_{\substack{c_a, c_b \in C \\ c_a \neq c_b}} G(i, j) P^{c_a}(i) P^{c_b}(j) = G(i, j) \sum_{c\in C} P^c(i)(1 - P^c(j)) \tag{3.16}$$

where G(i,j) is a Gaussian kernel filter, which is expressed as follows :

$$G(i, j) = \frac{1}{W} e^{-(\frac{\|D_i - D_j\|^2}{2\sigma_1^2} + \frac{\|I_i - I_j\|^2}{2\sigma_2^2})} \tag{3.17}$$

where $\frac{1}{W}$ is the normalized weight, D is the pixel spatial position of image I. σ_1 and σ_2 are the hyperparameters that control the scale of Gaussian kernels.

The final EN-Loss can be written as :

$$L_{en} = \sum_{i=0}^{N} \sum_{\substack{j=0 \\ j \neq i}}^{N} S(i)E(i, j) \tag{3.18}$$

According to what is described above, the implementation process and steps of our proposed LRM approach are shown in Table 1 by pseudo-code.

Table 1. Algorithm flow of our proposed approach.

```
Input: Images I with our image-level class labels $C_N$;
Output: The trained end-to-end network, LRM;

while iteration is true do
    Use the CAM to get the attention map as original seed;
    Binarize the attention map to generate the CAM-Mask;
    Perform mask erasing on the original image;
    Retrain the erased image using the AM-Loss
    to expand the foreground pixel of the seed region;
    Use (3.8) and (3.9) to get foreground probability $P_{fg}$ and background probability $P_{bg}$;
    Get the overall CAM probability map $P_{fg,bg}$ by combining $P_{fg}$ and $P_{bg}$;
    Calculate reliable CAM label $I_{cam}$ and CRF label $I_{crf}$ ;
    Get the reliable regions and label $I_f$ as pseudo labels from $I_{cam}$ and $I_{crf}$ using (3.11)(3.12);
    Train and update the whole network using loss function $L_{Joint} = L_{ce} + L_{en}$;
    Produce predictions;
end while
```

4. Experiments

4.1. Dataset description

4.1.1. Dataset

For the proposed dental disease segmentation work, the data comes from oral images of patients collected from a dental institution's long-term free clinic. The research object is oral images of patients aged between 18 and 70 taken using oral endoscopy A3S-X. With the help of dentists and experts, the acquired images were labeled by pixel-level classification using the Labelme tool. Among them, The red area is the lesion region of the decayed tooth, the green represents the lesion region of dental calculus, and the yellow is the lesion region of gingivitis. Our dental dataset includes four class annotations: three foreground objects and the background. In 3696 oral images, 2580 were used in the training set, 739 were kept in the validation set, and the remaining 377 were used as the test set. The dataset labeling process is shown in Figure 4. The label mask, output in this process, consists of three color-coded labeled images of the same size as the input oral images, showing the shapes of the three features of interest (decayed tooth, calculus, and gingivitis). It is worth noting that only the image-level classification labels are used in the training process to learn each feature of interest, while the manually labeled ground truth labels are used as the validation and test sets.

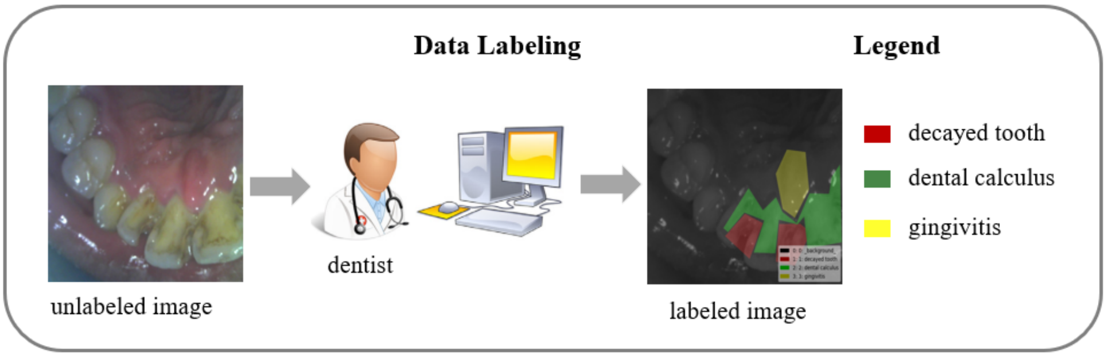


Figure 4. Illustration of the data labeling process. The ground truth labels with red for decayed tooth, green for dental calculus and yellow for gingivitis.

4.1.2. Evaluation metrics

In the evaluation phase, pixel-level labels manually labeled by experts are used as ground truth labels. The trained network takes an invisible test set and outputs a predicted label mask that contains any lesion region in the image for caries, dental calculus, and gingivitis. The segmentation performance is evaluated by comparing the MIoU score of the predicted label mask and the corresponding ground truth label mask.

4.2. Experimental settings

4.2.1. Image preprocessing

We resize the image to 448×768. For the problem of sparse and unbalanced datasets, we follow the work [32] and use automatic enhancement technology to increase the size and diversity of the dataset during training.

4.2.2. Implementation details

Baseline setting: We use ResNet [10] with 38 convolutional layers as the baseline network and fine-tune it to adapt to the current dental dataset. We remove all fully connected layers of the original network and perform dilated convolution for the last three resnet blocks (a resnet block is a set of residual units with the same output size), setting the dilated rate to 2 for the last third layer and 4 for the last two layers. For the semantic segmentation branch, we add two dilation convolution layers [33] of the same configuration after the backbone, with a kernel size of 3, a dilated rate of 12, and the padding size of 12. CE-Loss is computed for background and foreground individually. In the EN-Loss, σ_1 and σ_2 are set to 15 and 100 respectively.

Training strategy: We use Pytorch to implement our model. We train the network using SGD and terminate after 30 epochs with a batch size of 8. To address the tendency of early overfitting and instability training on the few-shot dataset, we use cosine [34] with warmup [35] to optimize SGD. The initial learning rate of warmup is set to 10^{-5}, and the maximum learning rate of 10^{-2} is reached after 2000 warmup iterations, followed by decaying with cosine, which can overcome the overfitting phenomenon and maintain the deep stability of the model. For the segmentation framework, we use DeepLab [36] to retrain the pseudo-pixel-level labels generated by the attention mining branch to finally achieve the segmentation prediction of dental lesions.

To generate reliable pseudo labels, the decay rate factor γ used to suppress background pixels in Eq (3.9) is set to 4 for $P_{fg,bg}$. The CRF parameter in Eq (3.10) follows the setting in [31]. In (3.11), when we select the high confidence region, the α value is set to 0.45. That is, 45% of the pixels are selected as the labeled pixels for each category. During validation and testing, dense CRF is used as a post-processing method to improve the accuracy of the segmentation boundaries, and the parameters are set as the default values given in [23]. During training, both two branches update the backbone network. During testing, only the segmentation branch is used to produce the predictions.

4.3. Experimental analysis and comparison

4.3.1. Comparison with previous approaches

Comparison with other networks dedicated to segmenting dental diseases. To highlight the strengths of our method, we compared it with the dental disease segmentation network based on improved U-Net [37]. Table 2 shows that our LRM method obtains a higher MIoU score, and the segmentation performance is improved by 11.18%. Some qualitative segmentation results are presented in Figure 5, which clearly shows that our LRM method (d) is closer to ground truth than the improved U-Net [37]. It also can be seen from Figure 5(I) that the segmentation network based on U-Net only identifies the most recognizable regions (c). In comparison, our LRM method captures the lesion area of the occlusal surface and the tiny lesion area inside the posterior alveolar teeth (d). This is attributed to the fact that our network shields regions of interest by CAM-Mask, forcing the network to focus on extracting more comprehensive features and semantic representations. Thus, our LRM method has a high capture ability for coarse-grained and fine-grained features.

Table 2. Segmentation performance comparisons with the dental disease segmentation network based on improved U-Net [37].

Method	baseline	dental calculus	decayed tooth	gingivitis	MIoU
Improved U-Net [37]	VGG16	62.68%	48.89%	43.42%	51.66%
Ours-LRM	VGG16	60.72%	64.68%	59.21%	61.54%
Ours-LRM	ResNet38	63.08%	64.34%	61.09%	62.84%

Furthermore, our segmentation results are compared with a CNN-based automated segmentation network for gingival disease [6]. Figure 6 shows that our segmentation results (d) cover more accurate target pixels. Especially in Figure 6(I), the automated segmentation network of gingival disease [6] carries a large number of error pixels in the tooth tissue area (c), which is because the label data limit their network during training. It is easy to learn a model with low generalization under the imbalance of the dataset, and the area with reddish color is mistakenly identified as gingivitis. While Our network learns in a weakly supervised way, explicitly modeling attention as part of training rather than passively using trained network attention offline. The advantage of this is that we can guide attention by using self-supervised and complementary learning attention mechanisms and generate reliable label masks, effectively improving the model's generalization ability.

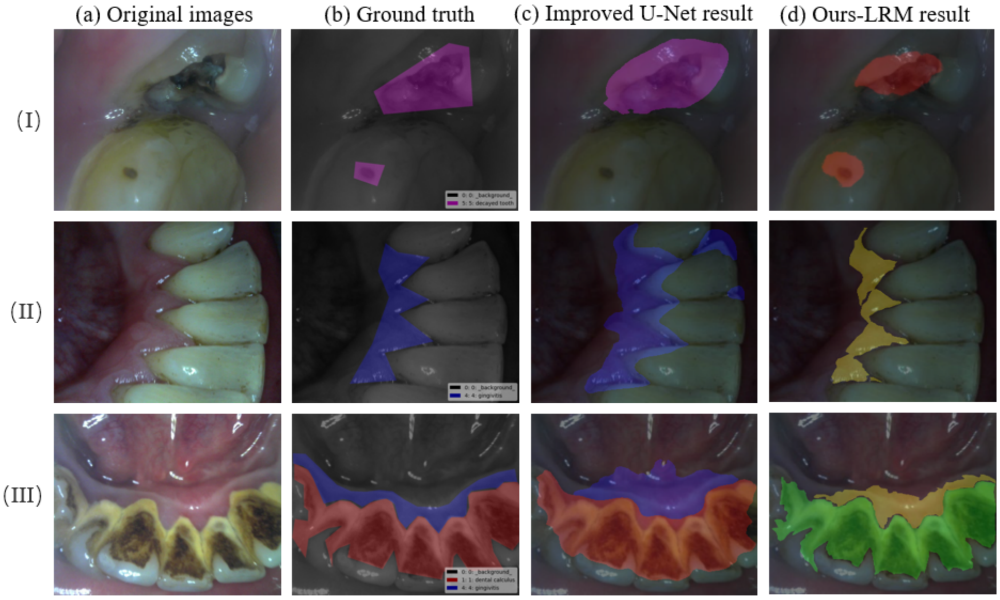


Figure 5. Dental disease segmentation results compared with the improved U-Net [37].

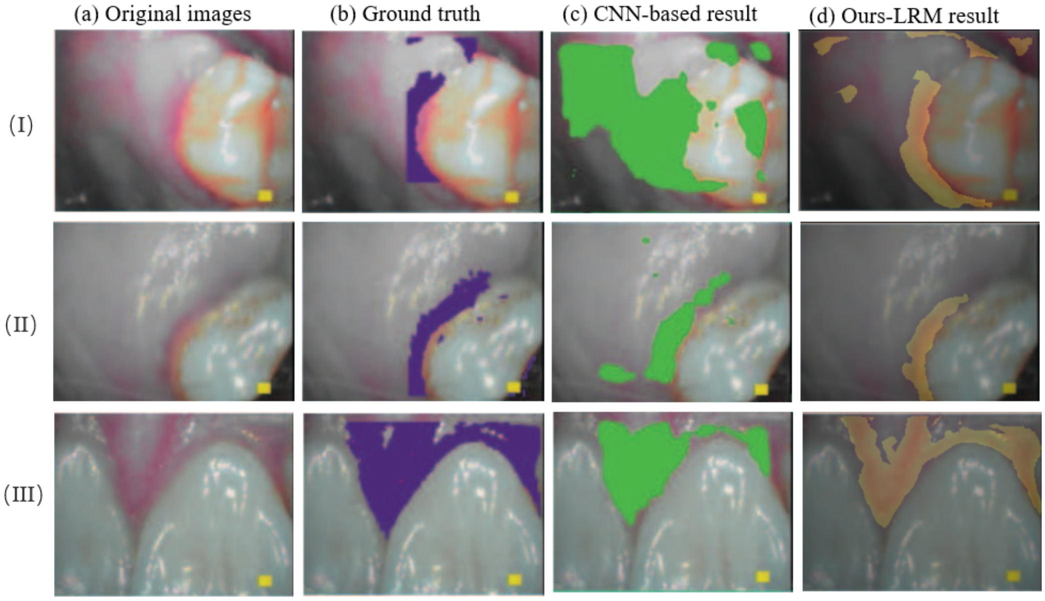


Figure 6. Gingival disease segmentation results compared with CNN-based network [6].

Comparison with other state-of-the-art weakly supervised semantic segmentation (WSSS) networks. Since there are no public datasets in the field of dentistry, research work is performed on private datasets. Establishing an objective comparison scale for research work between different datasets is difficult. Therefore, we feed our dataset into other WSSS networks (AdvErasing [21], SEAM [28], RCA [38]) with image-level labels for training and prediction. Table 3 shows the MIoU score of each class on the validation set. Compared to other weakly supervised semantic segmentation methods using different network structures, our LRM significantly presents a superior performance. The performance of our LRM based on ResNet38 [39] network is 2.71% higher than the SEAM [28] using the same baseline network and outperforms the RCA [38] using a deeper residual network. We also verify our LRM method on the VGG16 [40] network with a performance of 61.54%, which is 9.12% better than the AdvErasing [21] based on the VGG19 [40] network. It can be concluded that the performance improvement of our LRM method does not come from a deeper or larger network.

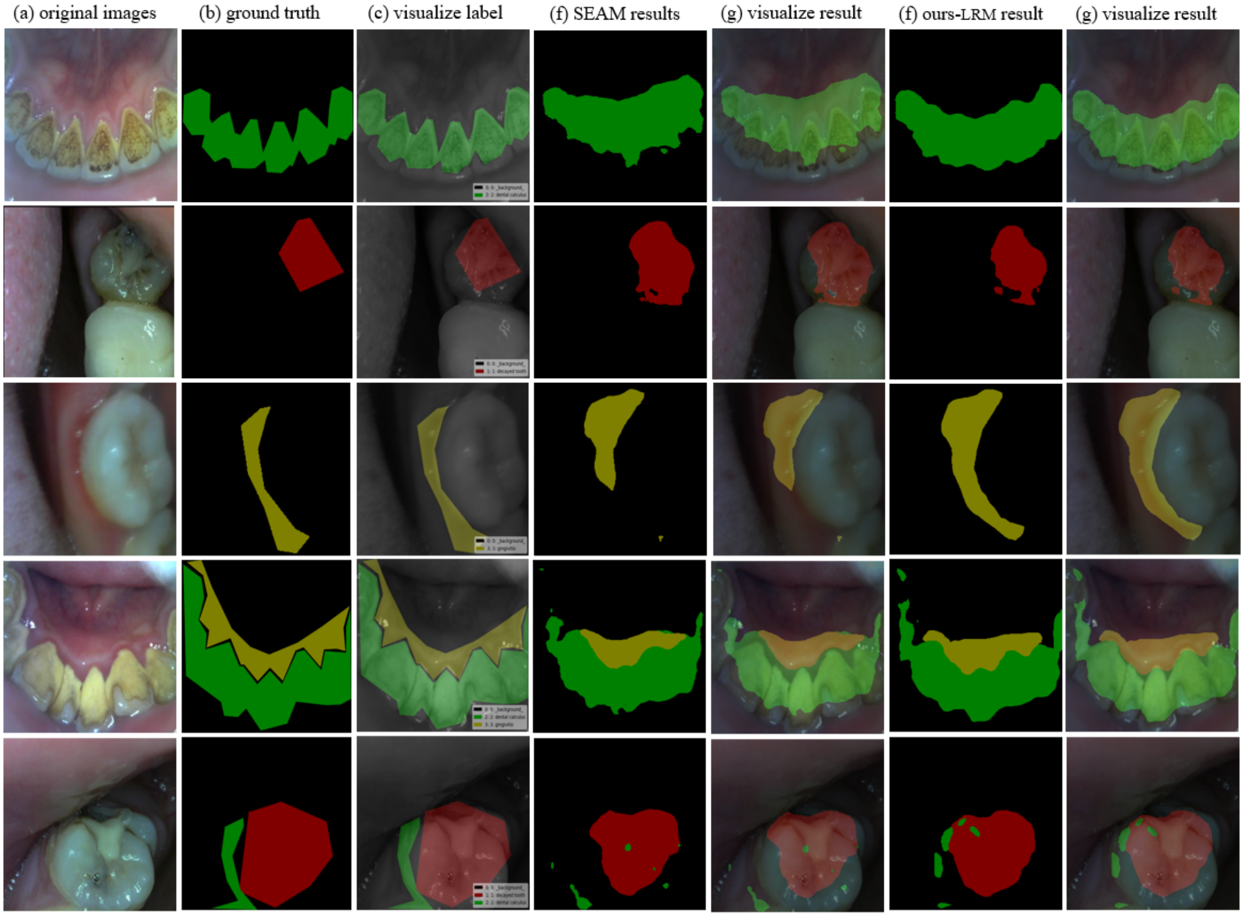


Figure 7. Qualitative segmentation results on validation set.

Considering most people's limited oral health knowledge, we further visualize segmentation results on the original image in combination with the segmentation mask, which can help people understand the segmentation results more intuitively. Figure 7 shows some qualitative results of semantic segmentation compared with the state-of-the-art SEAM [28]. We can observe that our LRM method help to discover more integral and accurate areas of the object. In addition, our segmentation result is more sensitive to the pixels at the boundary and presents a finer and smoother contour shape than the ground truth. However, it still carries little wrong backgrounds and foreground pixels compared to ground truth. In future research, we will attempt to inject a small amount of pixel-level supervision based on semi-supervised learning, closing the gap with the ground truth.

Table 3. Category performance comparisons with other state-of-the-art WSSS methods on validation set.

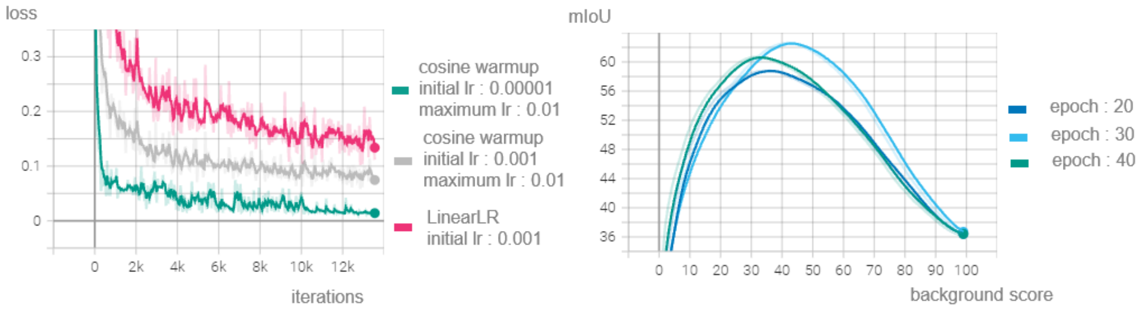
Method	Baseline	Dental calculus	Decayed tooth	Gingivitis	MIoU
AdvErasing [21]	VGG19	53.80%	54.64%	48.83%	52.42%
RCA [38]	ResNet50	57.20%	57.92%	52.45%	55.87%
SEAM [28]	ResNet38	61.28%	61.92%	57.18%	60.13%
Ours-LRM	VGG16	60.72%	64.68%	59.21%	61.54%
Ours-LRM	ResNet38	63.08%	64.34%	61.09%	62.84%

The above results show that our approach works well on single-label and multi-label tasks using only image-level labels. The reasons behind the performance improvement could be the combined effect of ACM and joint loss function instead of larger or deeper network structures. Specifically, more complete foreground regions are mined under the complementary learning mode of the ACM in the attention mining branch, which is because other finer feature regions can be mined by masking the most discriminative areas. The mined foreground regions are post-processed by CRF validation to effectively suppress the redundant background noise, generating better pseudo-pixel-level labels for the segmentation branch. The CE-Loss joined with the EN-Loss improves the segmentation performance in the semantic segmentation branch, which could be attributed to the fact that the Gaussian kernel filter in EN-Loss performs unlabeled or mislabeled pixels to alleviate the errors introduced by CE-Loss. Subsequently, we further confirmed this through ablation experiments in Section 4.3.2.

4.3.2. Ablation study

Our LRM consists of an attention mining branch and a segmentation branch. Pseudo-pixel-level labels with high confidence are generated using the attention compensation module (ACM) and CRF operations in the attention mining branch. The pixel-level segmentation prediction of dental lesions is performed by training and inferring through the joint loss function in the segmentation branch. We use ablation studies to illustrate their individual and joint effectiveness. In addition, we confirm the effect of different hyperparameter settings on segmentation performance by ablation experiments.

Hyperparameter: The process of model training is to minimize the loss function. During the training process, we set the batch size to 8 according to the number of samples and our hardware capacity. The model is tuned for the best performance by setting different hyperparameters. Figure 8(a) displays the results presented during the training process using different learning strategies. When the initial learning rate of warmup is set to 10^{-5}, and the maximum learning rate is set to 10^{-2}, the convergence is more stable, and the loss is minimized to 1.31%. Compared to the LinearLR adjustment strategy without the warmup, the loss is reduced by 13.53%. Because at the beginning of training, each data point is new to the model, and the model will quickly correct the data distribution. A large learning rate is likely to lead to overfitting of the data. After a period of training (2000 iterations), the model has some correct priors for the current mini-batch, and adopting a larger learning rate is less likely to skew the model. The distribution of the model has been relatively fixed in the later stage of training, and less new knowledge can be learned. We perform cosine decay learning after warmup to keep the stability of the model. It is concluded that the learning rate adjustment strategy with Cosine Warmup helps to slow down the early over-fitting of the mini-batch and maintain the deep stability of the model. Figure 8(b) shows the effect of three different epochs on segmentation performance. Performance is optimal when the epoch is 30. This is because the model has converged stably after 30 epochs of iteration. While too small an epoch is easy to underfit, too large is easy to overfit.



(a) (b)

Figure 8. Results of ablation experiments with hyperparameters.

In addition, Table 4 shows the effect of the regularization factor γ defined in Eq (3.9) and the choice of different background scores α in Eq (3.11). The γ in Eq (3.9) is set to 4, which can effectively suppress the background pixels and exert a positive regularizing effect. Setting 45% for α is a good choice to yield optimal performance. Since a smaller pseudo-mask size means that the region labeled for the segmentation branch has higher confidence, the segmentation network lacks sufficient information to learn, which will not achieve satisfactory performance. A larger size means that the labeled pixels have lower confidence, meaning that more incorrect labels are used, which corresponds to noise during training.

Table 4. Performance on our validation set based on different pseudo mask sizes. Ratio means the proportion of reliable regions which is mined by our method to the whole pixels.

γ				Ratio (α)			
(decay rate factor)	0.2	0.45	0.5	0.55	0.6	0.8	1.0
3	57.12%	59.37%	60.83%	59.28%	58.76%	50.91%	43.74%
4	57.90%	62.84%	62.41%	62.03%	59.52%	48.24%	41.33%
5	55.63%	57.21%	57.90%	58.40%	59.07%	51.28%	45.31%

Comparison with baseline: Table 5 gives an ablation study of each module in our method. It shows that the network using the attention compensation module (ACM) has a 9.17% improvement compared to the baseline. We also test baseline CAM with dense CRF operations to refine predictions. The results show that the CRF improves the MIoU score to 54.67%, which is lower than 58.90% using ACM. The generated pseudo labels can further improve the performance up to 62.84% on our train set after our LRM aggregating dense CRF as post-processing on the improved CAM using the attention compensation module (ACM).

Table 5. The ablation study for each part of LRM. **ACM**: attention compensation module.

baseline	ACM	CRF	MIoU
√			49.73%
√	√		58.90%
√		√	54.67%
√	√	√	**62.84%**

Loss function: In Table 6, we demonstrate the effectiveness of introducing joint loss by comparing the CE-Loss of baseline and joint loss of our LRM. The MIoU obtained using LRM with CE-loss is lower in the absence of joint loss. This result is because when only CE-Loss is considered, the mined reliable region with LRM cannot provide enough labels for segmentation model training, which will introduce some errors. The Gaussian kernel filter in EN-Loss mainly acts on unlabeled or mislabeled pixels to alleviate the errors introduced by CE-Loss. With the joint loss, the segmentation performance is improved substantially, from 56.07 to 62.84%, an improvement of 6.77%.

Table 6. Category performance comparisons with the baseline CE-Loss on validation set, where use top 45% pixels for α.

Loss	dental calculus	decayed tooth	gingivitis	MIoU
CE-Loss (Baseline)	56.43%	57.17%	54.63%	56.07%
Joint Loss (Ours)	63.08%	64.34%	61.09%	62.84%

Improved Localization Mechanism: We show qualitative results of attention maps generated by different methods in Figure 9, where LRM covers complete and accurate areas of the class of interest as well as fewer background areas around the class of interest compared with the baseline and SEAM. Obviously, the attention map generated by the baseline only contains local foreground information and carries some background noise (c). The attention map generated by SEAM has a little improvement, but it still cannot capture small target pixels (IV) and carry a small amount of background noise (d). In comparison, the attention map generated by our LRM method has less over-activation and complete activation coverage (e). Its shape is closer to the ground truth segmentation mask (b). This result is because the LRM network trains the entire network by minimizing the original image's classification loss while minimizing the class score of the image that obscures the object to be recognized during the training process. The trained network can better focus on the region to be recognized. Among them, the AM-Loss is used to compute the class activation score of the erased image, ensuring that the erased region does not contain a pixel that can activate the class. Under self-supervision of attention, the network will cover all important classification decision regions, expanding the foreground pixels. The CRF, a post-processing operation of the attention map, effectively suppresses excessive background noise.

Furthermore, Figure 9(IV) shows the decayed tooth of the lesion region in the posterior alveolar teeth and the occlusal surface. The attention maps of baseline and SEAM only located the lesion area of the occlusal surface without covering the target area inside the posterior alveolar teeth. However, the attention maps generated by our LRM can identify the lesion area on the occlusal surface and locate the tiny lesion area inside the posterior alveolar teeth. This result indicates that our LRM network model has strong generalization ability and robustness to dataset bias.

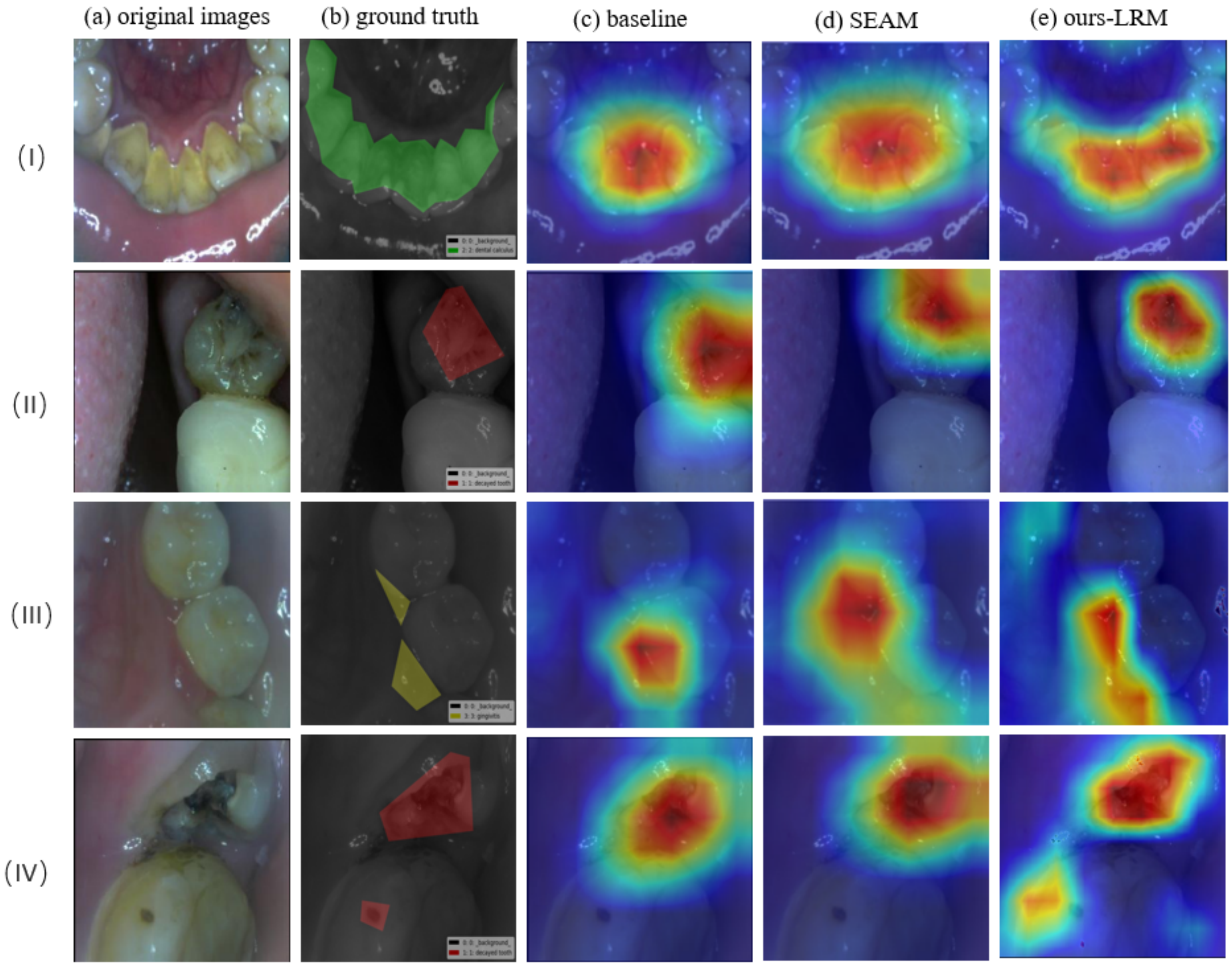


Figure 9. Qualitative results of attention maps.

5. Conclusions

We propose a framework for end-to-end semantic segmentation based on weakly supervised learning. The attention mining branch of the framework uses image-level annotations under an ACM and CRF operations to generate high-quality pseudo-pixel-level labels for the segmentation branch. The segmentation branch is trained by a joint loss function to predict pixel-level classification results. Our model effectively avoids the reliance on manual labeling and bridges the gap between pseudo labels and ground truth. It improves the robustness and generalization performance of the model in biased train sets. Applied to e-health systems and AI medical diagnosis systems, it solves the diagnostic challenges of patients in remote mountainous areas at the early stages of diseases. Moreover, it enables people to perform regular screening and early diagnosis of dental images taken by intraoral imaging devices through the system at home.

In AI dentistry, OCT has been widely used in dental diagnosis as a real-time and non-invasive tomography technology. To improve the generalizability of our method, we will subsequently attempt to make dental datasets based on OCT imaging and expand the scope of dental disease recognition of the proposed approach to achieve visually optimal lesion segmentation by superpixel segmentation.

Acknowledgments

This work was partially supported by the Natural Science Foundation of Shaanxi Province (No.2019JLM-10).

Conflict of interest

The authors declare there is no conflict of interest.

References

1. R. K. Meleppat, M. V. Matham, L. K. Seah, An efficient phase analysis-based wavenumber linearization scheme for swept source optical coherence tomography systems, *Laser Phys. Lett.*, **12** (2015), 055601. https://dx.doi.org/10.1088/1612-2011/12/5/055601
2. K. M. Ratheesh, L. K. Seah, V. M. Murukeshan, Spectral phase-based automatic calibration scheme for swept source-based optical coherence tomography systems, *Phys. Med. Biol.*, **61** (2016), 7652. https://dx.doi.org/10.1088/0031-9155/61/21/7652
3. R. K. Meleppat, C. Shearwood, L. K. Seah, M. V. Matham, Quantitative optical coherence microscopy for the in situ investigation of the biofilm, *J. Biomed. Opt.*, **21** (2016), 127002. https://doi.org/10.1117/1.JBO.21.12.127002
4. R. K. Meleppat, P. Prabhathan, S. L. Keey, M. V. Matham, Plasmon resonant silica-coated silver nanoplates as contrast agents for optical coherence tomography, *J. Biomed. Nanotechnol.*, **12** (2016), 1929–1937. https://doi.org/10.1166/jbn.2016.2297
5. W. J. Park, J. B. Park, History and application of artificial neural networks in dentistry, *Eur. J. Dent.*, **12** (2018), 594–601. https://doi.org/10.4103/ejd.ejd_325_18
6. A. Rana, G. Yauney, L. C. Wong, O. Gupta, A. Muftu, P. Shah, Automated segmentation of gingival diseases from oral images, in *2017 IEEE Healthcare Innovations and Point of Care Technologies (HI-POCT)*, (2017), 144–147. https://doi.org/10.1109/HIC.2017.8227605
7. X. Xu, C. Liu, Y. Zheng, 3D tooth segmentation and labeling using deep convolutional neural networks, *IEEE Trans. Vis. Comput. Graph.*, **25** (2019), 2336–2348. https://doi.org/10.1109/TVCG.2018.2839685
8. S. Sivagami, P. Chitra, G. S. R. Kailash, S. R. Muralidharan, UNet architecture based dental panoramic image segmentation, in *2020 International Conference on Wireless Communications Signal Processing and Networking (WiSPNET)*, (2020), 187–191. https://doi.org/10.1109/WiSPNET48689.2020.9198370

9. S. Li, Z. Pang, W. Song, Y. Guo, W. You, A. Hao, et al., Low-shot learning of automatic dental plaque segmentation based on local-to-global feature fusion, in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, (2020), 664–668. https://doi.org/10.1109/ISBI45749.2020.9098741

10. M. C. Kaya, G. B. Akar, Dental x-ray image segmentation using octave convolution neural network, in *2020 28th Signal Processing and Communications Applications Conference (SIU)*, (2020), 1–4. https://doi.org/10.1109/SIU49456.2020.9302495

11. W. Van Gansbeke, S. Vandenhende, S. Georgoulis, L. Van Gool, Unsupervised semantic segmentation by contrasting object mask proposals, in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, (2021), 10052–10062. https://doi.org/10.1109/ICCV48922.2021.00990

12. Z. Chen, T. Wang, X. Wu, X. S. Hua, H. Zhang, Q. Sun, Class re-activation maps for weakly-supervised semantic segmentation, in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2022), 959–968. https://doi.org/10.1109/CVPR52688.2022.00104

13. J. Ahn, S. Kwak, Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation, in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2018), 4981–4990. https://doi.org/10.1109/CVPR.2018.00523

14. J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2015), 3431–3440. https://doi.org/10.1109/CVPR.2015.7298965

15. J. Kim, J. Canny, Interpretable learning for self-driving cars by visualizing causal attention, in *2017 IEEE International Conference on Computer Vision (ICCV)*, (2017), 2961–2969. https://doi.org/10.1109/ICCV.2017.320

16. Z. Zhang, Y. Xie, F. Xing, M. McGough, L. Yang, MDNet: A semantically and visually interpretable medical image diagnosis network, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2017), 3549–3557. https://doi.org/10.1109/CVPR.2017.378

17. C. Song, Y. Huang, W. Ouyang, L. Wang, Box-driven class-wise region masking and filling rate guided loss for weakly supervised semantic segmentation, in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2019), 3131–3140. https://doi.org/10.1109/CVPR.2019.00325

18. M. Tang, F. Perazzi, A. Djelouah, I. Ben Ayed, C. Schroers, Y. Boykov, On regularized losses for weakly-supervised CNN segmentation, in *European Conference on Computer Vision*, (2018), 524–540. https://doi.org/10.1007/978-3-030-01270-0_31

19. K. K. Maninis, S. Caelles, J. Pont-Tuset, L. Van Gool, Deep extreme cut: From extreme points to object segmentation, in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2018), 616–625. https://doi.org/10.1109/CVPR.2018.00071

20. X. Zhang, Y. Wei, J. Feng, Y. Yang, T. S. Huang, Adversarial complementary learning for weakly supervised object localization, in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2018), 1325–1334. https://doi.org/10.1109/CVPR.2018.00144

21. Y. Wei, J. Feng, X. Liang, M. M. Cheng, Y. Zhao, S. Yan, Object region mining with adversarial erasing: A simple classification to semantic segmentation approach, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2017), 6488–6496. https://doi.org/10.1109/CVPR.2017.687

22. X. Wang, S. You, X. Li, H. Ma, Weakly-supervised semantic segmentation by iteratively mining common object features, in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2018), 1354–1362. https://doi.org/10.1109/CVPR.2018.00147

23. Z. Huang, X. Wang, J. Wang, W. Liu, J. Wang, Weakly-supervised semantic segmentation network with deep seeded region growing, in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2018), 7014–7023. https://doi.org/10.1109/CVPR.2018.00733

24. B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2016), 2921–2929. https://doi.org/10.1109/CVPR.2016.319

25. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in *2017 IEEE International Conference on Computer Vision (ICCV)*, (2017), 618–626. https://doi.org/10.1109/ICCV.2017.74

26. K. Baek, M. Lee, H. Shim, Psynet: Self-supervised approach to object localization using point symmetric transformation, in *Proceedings of the AAAI Conference on Artificial Intelligence*, **34** (2020), 10451–10459. https://doi.org/10.1609/aaai.v34i07.6615

27. F. Wan, P. Wei, J. Jiao, Z. Han, Q. Ye, Min-entropy latent model for weakly supervised object detection, in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2018), 1297–1306. https://doi.org/10.1109/CVPR.2018.00141

28. Y. Wang, J. Zhang, M. Kan, S. Shan, X. Chen, Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation, in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2020), 12272–12281. https://doi.org/10.1109/CVPR42600.2020.01229

29. P. Krähenbühl, V. Koltun, Parameter learning and convergent inference for dense random fields, in *Proceedings of the 30th International Conference on International Conference on Machine Learning*, (2013), 513–521.

30. K. Simonyan, A. Vedaldi, A. Zisserman, Deep in-side convolutional networks: Visualising image classification models and saliency maps, preprint, arXiv:1312.6034

31. T. Joy, A. Desmaison, T. Ajanthan, R. Bunel, M. Salzmann, P. Kohli, et al., Efficient relaxations for dense crfs with sparse higher-order potentials, *SIAM J. Imaging Sci.*, **12** (2019), 287–318. https://doi.org/10.1137/18M1178104

32. E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, Q. V. Le, Autoaugment: Learning augmentation policies from data, in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2019), 113–123. https://doi.org/10.1109/CVPR.2019.00020

33. Z. Wu, C. Shen, A. Van Den Hengel, Wider or deeper: Revisiting the resnet model for visual recognition, *Pattern Recognit.*, **90** (2019), 119–133. https://doi.org/10.1016/j.patcog.2019.01.006

34. P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, et al., Accurate, large minibatch SGD: Training imageNet in 1 hour, preprint, arXiv.1706.02677

35. T. He, Z. Zhang, H. Zhang, Z. Zhang, J. Xie, M. Li, Bag of tricks for image classification with convolutional neural networks, in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2019), 558–567. https://doi.org/10.1109/CVPR.2019.00065

36. L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A. L. Yuille, Semantic image segmentation with deep convolutional nets and fully connected CRFs, preprint, arXiv.1412.7062

37. W. Shang, Z. Li, Y. Li, Identification of common oral disease lesions based on U-Net. in *2021 IEEE 3rd International Conference on Frontiers Technology of Information and Computer (ICFTIC)*, (2021), 194–200. https://doi.org/10.1109/ICFTIC54370.2021.9647420

38. T. Zhou, M. Zhang, J. Li, Regional semantic contrast and aggregation for weakly supervised semantic segmentation. in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2022), 4289–4299. https://doi.org/10.1109/CVPR52688.2022.00426

39. Z. Wu, C. Shen, A. Van Den Hengel, Wider or deeper: Revisiting the resnet model for visual recognition, *Pattern Recognit.*, **90** (2019), 119–133. https://doi.org/10.1016/j.patcog.2019.01.006

40. K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, preprint, arXiv.1409.1556