IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 6 October 2023, accepted 30 October 2023, date of publication 2 November 2023, date of current version 15 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3329975

RESEARCH ARTICLE

Cracked Tongue Extraction Model Based on Improved U-Net Method

ZIHAO ZHANG[1], JIANHUA ZHENG[1,2], RUOLIN ZHAO[1], SHUANGYIN LIU[1], ZHENGJIE LIU[3,4], AND JINHE WANG[5]

[1]College of Information Science and Technology, Zhongkai University of Agriculture and Engineering, Guangzhou 510225, China
[2]Guangdong Provincial Key Laboratory of Traditional Chinese Medicine Informatization, Guangzhou 510630, China
[3]Guangdong Provincial Hospital of Chinese Medicine, Guangzhou 510120, China
[4]The Second Affiliated Hospital of Guangzhou University of Chinese Medicine, Guangzhou 510120, China
[5]Xiyuan Hospital of China Academy of Chinese Medical Sciences, Beijing 100091, China

Corresponding author: Jianhua Zheng (zhengjianhua@zhku.edu.cn)

This work was supported in part by the Research Fund Program of Guangdong Provincial Key Laboratory of Traditional Chinese Medicine Informatization under Grant 2021B1212040007 and Grant 2021503; in part by the National Key Research and Development Program of China under Grant 2018YFC2002500; in part by the Natural Science Foundation of Guangdong Province under Grant 2022B1515120059 and Grant 2023A1515011230; in part by the National Natural Science Foundation of China under Grant 61871475; in part by the Innovation Team Project of Universities in Guangdong Province under Grant 2021KCXTD019; in part by the Science and Technology Planning Project of Yunfu under Grant 2023020202, Grant 2023020203, and Grant 2023020205; in part by the Science and Technology Program of Guangzhou under Grant 2023E04J1238, Grant 2023E04J1239, Grant 2023E04J0037, and Grant 2023E04J0037; in part by the Guangdong Science and Technology Project under Grant 2020B0202080002; in part by the Guangdong Province Graduate Education Innovation Program Project under Grant 2022XSLT056 and Grant 2022JGXM115; in part by the Major Science and Technology Special Projects in Xinjiang Uygur Autonomous Region under Grant 2022A02011; in part by the Meat Pigeon Industrial Park Technology Research and Development Project in Xingning, Meizhou (Construction and Promotion of Visual Information Platform) under Grant GDYNMZ20220527; and in part by the Science and Technology Planning Project of Heyuan under Grant 202305.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Ethics Committee of Guangdong Provincial Hospital of Chinese Medicine.

ABSTRACT Tongue diagnosis holds significant importance in Traditional Chinese Medicine (TCM), with cracked tongues serving as a key diagnostic feature. However, the considerable variability in the morphology, depth, and distribution of tongue cracks poses a challenge for accurate extraction. In this paper, a novel deep learning approach is proposed to enhance the decoder of the U-Net model for cracked tongue extraction by incorporating the Hybrid Parallel Attention Mechanism (HPAM). The inclusion of HPAM enables the model to better concentrate on the small-scale feature information of tongue cracks, thereby improving the accuracy of crack segmentation. Experimental results demonstrate the effectiveness of the proposed method across all three tongue crack datasets. The method achieves a MIoU of 69.31% on the open environment dataset, 76.05% MIoU on the non-open environment dataset, and an overall MIoU of 76.92% on the combined dataset. These results signify a significant improvement over existing methods. This study not only offers an effective approach for automating the extraction of cracked tongues but also contributes to the automation and accuracy of tongue diagnosis, thereby benefiting the field of TCM.

INDEX TERMS Cracked tongue, deep learning, attention mechanism.

I. INTRODUCTION

Cracked Tongue, also known as tongue fissures, refers to the notable depressions and elevations present on the surface of the tongue. This distinctive morphology holds significant importance in facial recognition and disease detection [1], [2]. Apart from its widespread application in traditional Chinese medicine diagnosis [3], recent research has increasingly shown that cracked tongue can reflect changes in overall health and serve as an indicator for early disease detection [2]. Hence, the efficient and accurate extraction of cracked tongue information has become a crucial concern.

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

Previously, some researchers employed traditional methods for extracting images of fissured tongues. Li et al. used a hyperspectral tongue imaging device to capture images of the tongue and then applied a classification algorithm based on Hidden Markov Models, achieving a certain level of effectiveness [4]. With the advancement of deep learning techniques, Convolutional Neural Networks (CNNs) have become the mainstream approach in medical image analysis. Numerous researchers have applied CNNs to the domain of tongue feature extraction and recognition, yielding promising results [5], [6], [7]. For example, Huang et al. applied deep learning methods to construct a tongue segmentation model for the segmentation of mobile-acquired tongue images in open and complex environments [8]. Ruan et al. constructed an efficient tongue image segmentation model by optimizing the UNet network and designed a new network to specifically handle tongue edge segmentation [9]. Song et al. proposed RAFF-NET for tongue region segmentation [10]. The above-mentioned study achieved good results, but did not investigate tongue fissures. Existing methods that combine deep learning and tongue diagnosis primarily fall into two categories: object detection [11] and instance segmentation. For instance, Hui et al. propose a weakly supervised method for training the tooth-mark and crack detection model by leveraging fully bounding-box level annotated and coarse image-level annotated tongue images, achieving an accuracy of 0.865 in cleft palate recognition [12]. Object detection methods have shown certain effectiveness in tongue diagnosis; however, they fail to obtain precise contour lines for cracked tongue. As a result, some researchers have employed segmentation methods to extract the contour of cracked tongue. For example, Xue et al. used crack and non-crack regions to train AlexNet, extracting deep features of the crack region, and finally performed classification using Support Vector Machines (SVM) [9]. Yan et al. proposed the Segmentation-Based Deep Learning (SBDL) model for cracked tongue image extraction and recognition [10]. Li et al. improved the partial encoder of the Unet architecture by introducing a global convolutional network module to address the encoder's inability to extract relatively abstract high-level semantic features, thereby achieving cracked tongue extraction. However, this model only achieved a MIoU score of 0.473 on the test set [11]. Although Transformer-based models have achieved excellent performance in some application scenarios, they require large training datasets [13], which leads to suboptimal results when dealing with small datasets. Moreover, the extraction of cracked tongue is susceptible to environmental interference, making it challenging to extract all the fissures accurately.

Considering the intricacy of the backdrop in the patient's uploaded tongue photographs, in order to enhance the precision of tongue fissure extraction, this paper devises a Hybrid Parallel Attention Mechanism, augmenting the U-net framework. The primary objective is to accomplish tongue fissure extraction. Firstly, to address the issue of indistinct differentiation between the foreground and background of tongue fissures, this paper reconfigures the U-net architecture and introduces the HPAM (Hybrid Parallel Attention Mechanism) module to amplify the model's capacity for capturing intricate details of tongue fissures and intensifying its focus on the fissure regions. This refinement aims to elevate the accuracy of segmentation. Secondly, data augmentation techniques and regularization strategies are employed to combat the issue of overfitting that often arises when training on small-scale datasets. Concurrently, the proposed model demonstrates a MIoU (Mean Intersection over Union) value of 76.92% on the test set, signifying high accuracy and robustness in the task of tongue fissure extraction. This research provides valuable insights for subsequent investigations.

The principal contributions of this paper are outlined as follows:

(1) In the realm of tongue fissure segmentation, this study proposes an enhanced U-net-based algorithm for tongue fissure extraction, effectively overcoming the difficulties and challenges associated with this task.
(2) A hybrid parallel attention mechanism strategy is introduced to facilitate the model in placing greater emphasis on the fissures themselves, thereby reducing interference stemming from complex background and environmental factors present in the tongue fissure images. Consequently, the performance of the model in tongue fissure extraction is substantially improved.
(3) A dataset of tongue fissures is generated. To address overfitting concerns when working with this limited dataset, data augmentation techniques and regularization strategies are employed, significantly bolstering the model's generalization capabilities.

The structure of this paper is organized as follows: The first section introduces the background and relevant research pertaining to this study, as well as the key contributions made. The second section delineates the data sources and pre-processing methodologies employed, along with the proposed methodology. The third section verifies the efficacy of the proposed method in tongue fissure extraction through experimentation and comparative analysis with alternative approaches. The fourth section discusses the findings of this study and presents future research prospects. Finally, the fifth section concludes this research endeavor.

II. MATERIALS AND METHODS

A. DATA SETS AND PREPROCESSING

1) DATA SOURCES

In this paper, we collected 132 images of tongues exhibiting cracks using web crawlers and color images from TCM books [27]. Additionally, we included 200 tongue photos obtained from Guangdong University of Traditional Chinese Medicine, resulting in a comprehensive dataset of 332 images. This dataset was subsequently divided into two subsets: 155 images captured in an open environment and

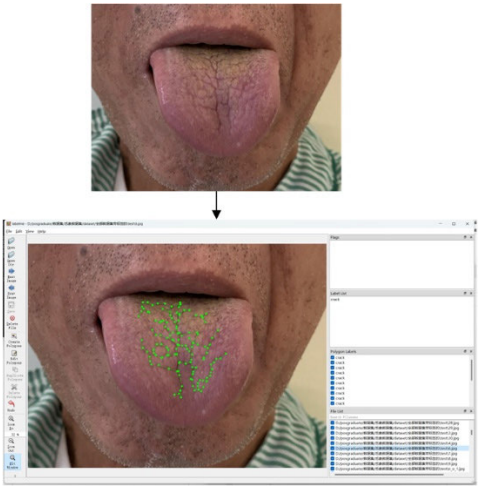

FIGURE 1. Schematic diagram of dataset labeling.

177 images captured in a non-open environment. The ''open environment'' subset comprises images that encompass the tongue, along with partial depictions of the head and body, providing visibility of the surroundings. In contrast, the ''non-open environment'' subset includes images solely focused on the cracked tongue, with minimal inclusion of background elements, thus mitigating interference from extraneous factors.

2) DATA LABELING AND ANALYSIS

The dataset utilized in this paper adheres to the Pascal VOC2007 standard format, with the image labeling tool Labelimg employed for the annotation process. The labeling procedure using this software is visually illustrated in Figure 1. Specifically, all tongue cracks in the image are depicted with a polygon box marked as a crack and a multifold segment.

In the process of tongue crack detection and extraction, the accurate extraction of tongue cracks is challenging due to several factors. These include the large number of tongue cracks, the minimal contrast between the cracks and the background, as well as variations in environmental conditions and filming equipment. For instance, as shown in Figure 2a, the captured image may be affected by varying light intensities, resulting in a darker appearance of the tongue surface. In Figure 2b, the image is influenced by the shooting distance, leading to a smaller tongue target area, albeit with fewer tongue cracks. Figure 2c illustrates an image captured at a close distance with shallow cracks, resulting in limited differentiation between the cracks and the background. Figure 2d demonstrates a scenario where the tongue exhibits a high number of complex cracks. Achieving pixel-level crack extraction using target detection methods becomes challenging in such cases. Therefore, this paper utilizes image segmentation techniques to accomplish crack extraction. The aforementioned characteristics of the image data intensify the challenges associated with tongue crack extraction.

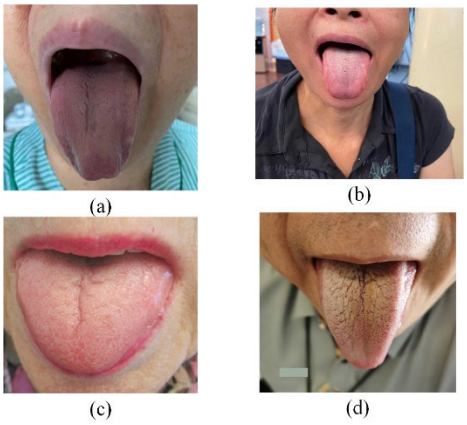

FIGURE 2. Challenges encountered in the tongue crack extraction task.

3) DATA ENHANCEMENT

Because the original dataset is small, consisting of only 332 cracked tongue images, a dataset of this size can lead to poor model generalization performance. Therefore, this paper randomly selects approximately 30% of the data from both the open environment dataset and the non-open environment dataset as the test set. Before doing so, six data enhancement methods are applied to augment the remaining images. These methods include image flip (Figure 3a), random rotation (Figure 3b), contrast enhancement (Figure 3c), random color dithering (Figure 3d), brightness enhancement (Figure 3e), and color enhancement (Figure 3f). The enhanced data were considered to be filtered and labeled, and after removing some images that could not be recognized by the naked eye due to texture loss caused by overexposure, there were 759 enhanced images in the open environment dataset and 720 enhanced images in the non-open environment dataset. The two datasets are divided into training and validation sets in the ratio of 8:2. By undergoing the data enhancement process, the new images acquire more comprehensive image features, which in turn enhance the training model's performance and yield improved results.

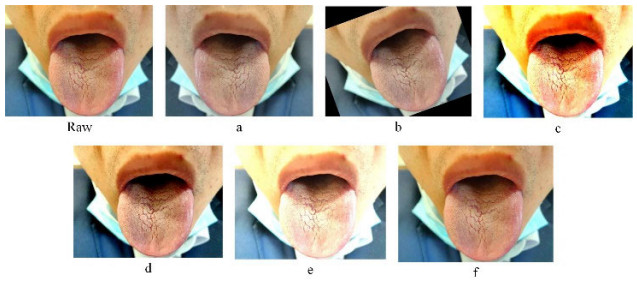

FIGURE 3. Image enhancement methods.

B. MODEL STRUCTURE DESIGN AND PRINCIPLES

1) U-NET MODEL

The U-net derives its name from its network structure resembling the shape of the letter ''U.'' It is a convolutional neural

network model specifically designed for image segmentation tasks [14]. It demonstrates remarkable suitability for medical image segmentation, particularly when accurate segmentation of smaller targets such as cells, blood vessels, and the like is required. As illustrated in Figure 4, the U-Net model bears resemblance to a self-encoder, encompassing both a downsampling path and an upsampling path. The downsampling path comprises a convolutional block and a downsampling block, enabling the extraction of global features from the input image. Meanwhile, the upsampling path facilitates the restoration of spatial information in the segmentation output. During the training phase, the U-Net employs a technique called jump-join, effectively connecting the feature maps from the downsampling path to their corresponding counterparts in the upsampling path, thereby achieving precise and detailed segmentation outcomes. The versatility of the U-Net model extends beyond medical image segmentation, holding promising applications in various other domains requiring image segmentation. Its successful deployment also yields valuable insights for image segmentation tasks in diverse fields.

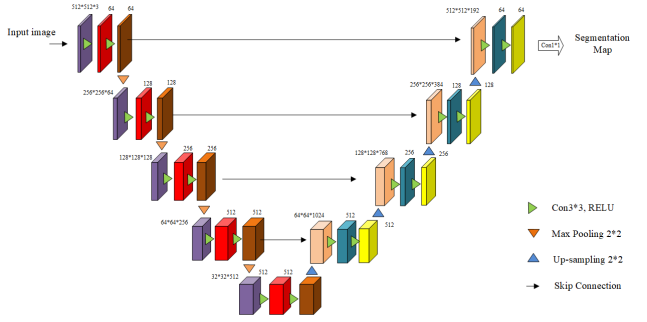

FIGURE 4. U-net model structure.

2) HYBRID PARALLEL ATTENTION MECHANISM

To focus on tongue fissure information in tongue images, a Hybrid Parallel Attention Mechanism (HPAM) is designed in this paper, which is computed by three different attention modules in parallel on three tracks, and finally the output of the three modules is summed up. The Hybrid Parallel Attention Mechanism (HPAM) consists of three main parallel modules: the SENet module [15], the SAM [16] module, and the CAM module [16]. Given a feature map input, where, and denote the height, width, and number of channels of the feature map, respectively, each of these three modules processes the input feature map to generate a new feature map.

SENet module: The SENet module first obtains the channel descriptor through a global averaging pooling operation and then implements the rescaling of the channel through two fully connected layers.

Assume that the weights of the two fully connected layers in the SENet module are $W_1 \in \mathbb{R}^{C/r \times C}$ and $W_2 \in \mathbb{R}^{C \times C/r}$, where r is the ratio of the number of channels reduced, then the output of the SENet module can be expressed as:

$$X_{\text{SENet}} = F_{\text{scale}}\left(X, W_2\delta\left(W_1 F_{\text{avg}}(X)\right)\right) \cdot X \tag{1}$$

where F_{scale} stands for the channel rescaling (or scale) function, F_{avg} denotes the global average pooling, δ denotes the ReLU activation function, and W_1 and W_2 are the weights of the two fully connected layers.

SAM module: The spatial attention mechanism (SAM) obtains the spatial attention map by calculating the maximum and average values of the input feature map in the channel dimension. The attention map is then activated using a sigmoid function. the output of the SAM module can be expressed as:

$$X_{SAM} = \sigma\left(F_{\max}(X) + F_{\text{avg}}(X)\right) \cdot X \tag{2}$$

where σ represents the sigmoid activation function, F_{max} denotes the maximum operation in the channel dimension, and F_{avg} denotes the global average pooling.

CAM module: The channel attention mechanism (CAM) begins by computing the maximum and average values of the input feature map in the spatial dimension. This calculation results in a channel attention map, which is subsequently activated using a sigmoid function:

$$X_{CAM} = \sigma\left(F_{\max}\left(X'\right) + F_{\text{avg}}\left(X'\right)\right) \cdot X \tag{3}$$

where σ symbolizes the sigmoid activation function, F_{max} denotes the maximum operation in the spatial dimension, F_{avg} denotes the average pooling in the spatial dimension, and X' is the result of global pooling of X in the spatial dimension, i.e., $X' = GlobalPool(X)$.

After these three modules have processed the input feature maps, we will get three new feature maps, denoted as X_{SENet}, X_{SAM} and X_{CAM}. Then, we add these three feature maps by value to obtain the output feature map X_{HPAM} of HPAM:

$$X_{HPAM} = X_{SENet} + X_{SAM} + X_{CAM} \tag{4}$$

The structure diagram of HPAM is shown in Fig. 5. By introducing SENets, it can make the network more sensitive to such fine features as tongue crack, and thus improve the accuracy of crack segmentation. The use of Spatial Attention can make the network pay more attention to the region closely related to the tongue crack, thus avoiding some misclassification problems. The introduction of Channel Attention allows the network to adjust the weights adaptively between different channels, thus enhancing the discriminative ability of tongue crack and further improving the segmentation accuracy. The experiments in Section III-F of this paper also verify this idea.

3) HAU-NET MODEL

The U-Net model has found extensive applications in medical image segmentation. However, one limitation is that the decoder component may struggle to accurately reconstruct detailed information. This is due to the loss of spatial information during the downsampling process, making it challenging

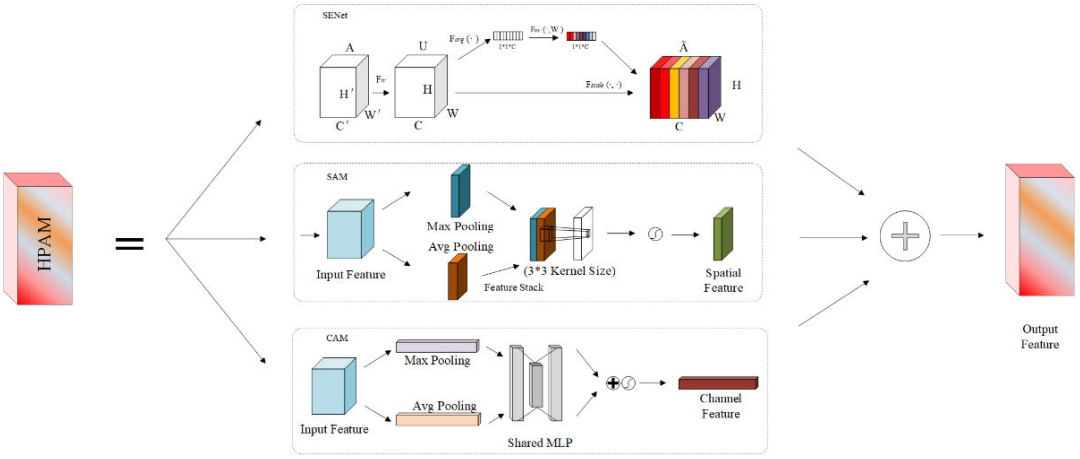

FIGURE 5. HPAM structure.

for the decoder to recover precise details such as edges and contours. This issue becomes particularly prominent in the context of tongue crack extraction, where distinguishing crack edges from the similarly colored tongue body proves difficult.

To address this challenge, this paper proposes the HAU-net, which integrates the Hybrid Parallel Attention Mechanism (HPAM) into the U-Net decoder. As depicted in Figure 6, three HPAM modules are incorporated into the decoder network after the upsampling stage in U-Net. This allows for the fusion of features from different scales in the encoder and extraction of more informative features during the decoder stage using multi-track parallel attention modules.

The HPAM module is not embedded in the encoder because the downsampling operation of the U-Net encoder leads to a loss of detailed information in the original tongue crack image, resulting in poor recognition performance, as demonstrated in the experimental section of Section III in this paper. By incorporating the HPAM modules in the decoder, the HAU-net model can focus more on the relevant tongue crack information within the mixed feature maps of different scales. Consequently, this approach effectively enhances the accuracy and efficiency of the model in detecting and identifying tongue cracks.

C. LOSS FUNCTION

The function utilized in the tongue crack extraction model consists of two components: the cross-entropy loss and the Dice loss. This can be represented by Equation 5.

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{BCE}} + \mathcal{L}_{\text{Dice}} \tag{5}$$

$$\mathcal{L}_{\text{BCE}} = -\sum_{i} \left(t_i \ln\left(\hat{t}_i\right) + (1 - t_i) \ln\left(1 - \hat{t}_i\right)\right) \tag{6}$$

where t_i and $\hat{t}_i$ denote the cracked region of the tongue predicted by the network and the cracked region of the true value, respectively. To deal with the class imbalance problem, this paper also uses the dice loss [12], which is defined as follows:

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2 \cdot \left\langle t_{(h,w)}, \hat{t}_{(h,w)} \right\rangle + \sigma}{\left\| t_{(h,w)} \right\|_1 + \left\| \hat{t}_{(h,w)} \right\|_1 + \sigma} \tag{7}$$

where $_{(h,w)}$ is the pixel coordinate, and σ is the Laplacian smoothing factor that accelerates the convergence rate of the network. Here, we will set σ to 1e-5 in our work.

TABLE 1. Experimental environment configuration parameters.

Parameters	Value
Input size	512×512
Batch size	4
Optimizer	Adam
Momentum	0.9
Initial learning rate	1e4
Epoch	300

III. EXPERIMENTAL RESULTS AND ANALYSIS

A. EXPERIMENTAL ENVIRONMENT AND CONFIGURATION

The test environment in this paper is: intel®Core™i5-6500Q CPU @2.30GHz, GeForce GTX 3090 (24GB) graphics card, Ubuntu 18 OS, python 3.8.

The parameters trained in this paper are listed in the following table:

B. TRAINING PROCESS

The experimental training process is illustrated in Figure 7. Initially, the dataset was divided into a test set comprising 30% of the images, while the remaining images were further partitioned into an 80% training set and a 20% validation set after pre-processing and data augmentation. The training and validation sets were utilized for model training, hyperparameter tuning within each epoch, and optimization method adjustment, aiming to achieve optimal results for tongue crack extraction.

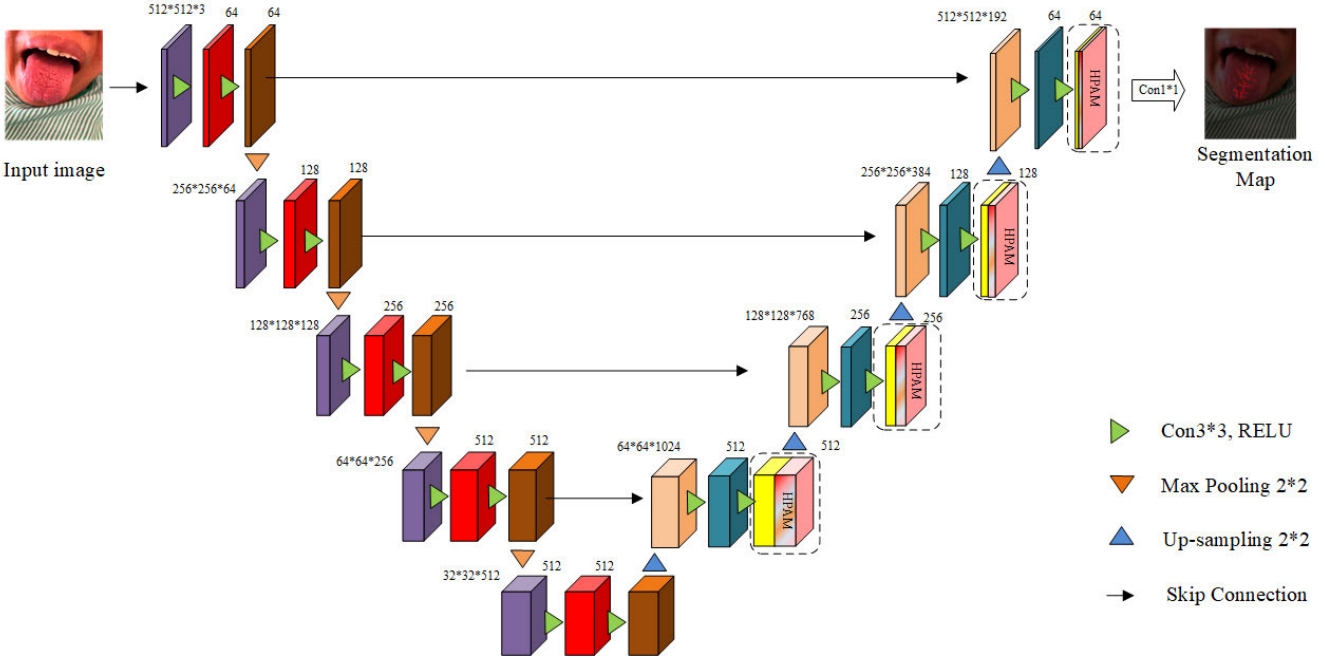

FIGURE 6. HAU-net model structure.

For all models, an input size of 512 × 512 was utilized, with a batch size of 4. The Adam optimizer was employed, with an initial learning rate of 1e-4 and momentum set to 0.9. The learning rate was adjusted using the cosine annealing algorithm.

To assess the model's superiority, a comparison was made with mainstream segmentation algorithms, including U-net_vgg16 with vgg16 as the backbone network, U-net_resnet50 with resnet50 as the backbone network, U-net++ [17], deeplabV3 [18], Segnet [19], and FRCnet [20]. Initially, each model exhibited a relatively high loss. However, after 300 iterations of network training, all seven models displayed a converging trend in both the training and validation sets, reaching their lowest loss points, as depicted in Figure 8. The figure demonstrates that U-net_resnet50 achieves fast convergence with a training set loss of approximately 0.2. However, its performance in the validation set is less favorable, oscillating around 0.35. On the other hand, the U-net model enhanced with the HPAM module, referred to as HAU-net, exhibits the lowest loss convergence in the validation set. This outcome indicates that HPAM enhances the generalization performance of the original model.

C. EVALUATION INDICATORS

In order to verify the performance of the model for tongue crack extraction, MIoU, Recall, Precision, Accuracy, and Dice coefficients are used as evaluation indexes for model segmentation performance in this paper. The calculation equations are as follows:

$$\text{Acc} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{8}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{9}$$

$$\text{IoU} = \frac{\text{TP}}{\text{FN} + \text{FP} + \text{TP}} \tag{10}$$

$$\text{MIoU} = \frac{1}{N}\sum_{i=1}^{N} \text{IoU} \tag{11}$$

$$\text{Dice} = \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FP} + \text{FN}} \tag{12}$$

In the equation: (True Positive, TP) is the number of correct results judged as correct; False Positive,FP is the number of correct results judged as wrong; True Negative TN is the number of wrong results judged as wrong; False Negative,FN is the number of wrong results judged as wrong. False Negative (FN) is the number of incorrect results judged as correct.

IoU is a commonly used evaluation metric to measure the degree of overlap between the predicted segmentation results and the true segmentation results, which is defined as the ratio of the intersection area of the predicted segmentation results to the merged area of the true segmentation results. It is a metric used to measure the average IoU of the model over multiple categories. It is the average of IoU for each category, as shown in Equation 10, where C denotes the number of categories and IoU_i denotes the IoU of i categories.

D. PERFORMANCE COMPARISON OF THE MODELS BEFORE AND AFTER DATA AUGMENTATION

To discuss the impact of data augmentation methods in data preprocessing on model performance, this study conducted comparative experiments on U-net_vgg16, HAU-net, U-net++, DeeplabV3, Segnet, and FRCnet before and after

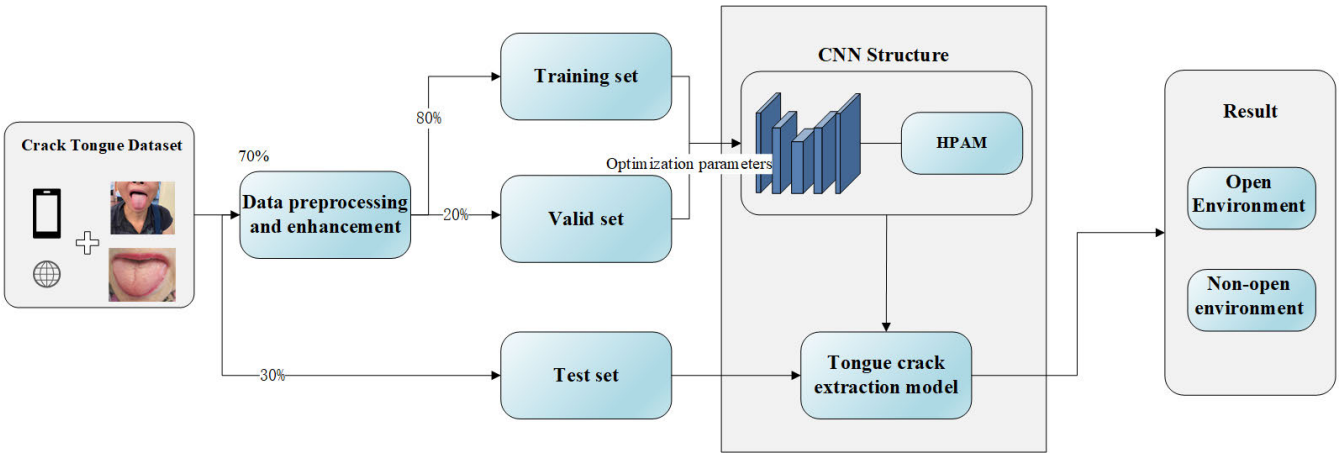

FIGURE 7. Training flow chart.

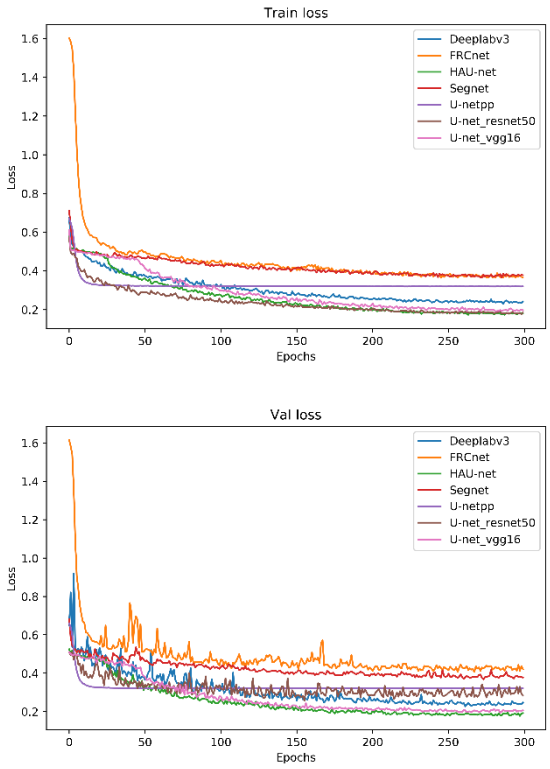

FIGURE 8. Training loss graph.

data augmentation on the test dataset, as shown in Table 2. It can be observed that all models show varying degrees of performance improvement after using data augmentation in preprocessing. The most significant improvement is observed in the U-net++ model, with a 24.75 increase in the MIoU metric. Augmented models exhibit better robustness when facing changes in illumination, noise, or shooting angles in input data. Data augmentation methods assist the model in adapting to these variations, thereby enhancing the model's robustness. By introducing techniques such as random rotation and random color jitter, the model can learn more diverse patterns, making it more stable for the task of tongue crack extraction under different conditions.

TABLE 2. Performance of the models before and after data augmentation.

Models	MIoU	MIoU_DE
U-net_vgg16	57.74	75.07
U-net++	49.55	74.3
DeeplabV3	56.24	59.06
Segnet	50.12	58.45
FRCnet	54.52	58.54
HAU-net	58.63	**76.92**

E. PERFORMANCE COMPARISON OF DIFFERENT MODELS UNDER DIFFERENT DATA SETS

To verify the performance of the model in this paper on the overall dataset containing all images, the open environment dataset, and the non-open environment dataset, HAU-net is tested against U-net, U-net++, DeeplabV3, Segnet, and FRCnet on the test set in this paper, respectively.

1) RESULTS ON THE OVERALL DATASET

The experimental results are shown in Table 3, and the results indicate that the HAU-net model proposed in this paper shows the best performance in all these metrics. Specifically, it achieves 76.92 on MIoU, which is significantly better than other models. Similarly, on the Dice coefficient, HAU-net reaches 0.810, which is the highest among all models. On the Recall metric, HAU-net also outperforms the other models (0.847) and is second only to the U-net++ model. Finally, HAU-net achieves 97.76 in Accuracy, a score that is the highest among all models, including Hausdorff Distance(HD), which measures the dissimilarity between predicted and ground truth masks.

U-net_Vgg16 and Resnet50 also perform quite well when used as a backbone, especially on MIoU, reaching 75.19 and

75.07, respectively (based on Resnet50). This is the reason why we chose to improve based on U-net in this paper. Other models, such as U-net++, DeeplabV3, Segnet, and FRCnet, are competitive in some metrics, but in general, their performance has yet to be improved compared to the HAU-net model.

TABLE 3. Performance of different models on the overall dataset.

Models	Backbone	MIoU	Recall%	Dice	HD	Accuracy%
U-net	Vgg16	75.19	85.89	0.801	6.53	99.74
U-net	Resnet50	75.07	83.76	0.807	6.38	99.73
U-net++		74.3	**90.3**	0.788	7.80	99.68
DeeplabV3		59.06	68.91	0.643	7.18	99.39
Segnet		58.45	59.92	0.654	7.91	99.65
FRCnet		58.54	64.42	0.634	7.43	99.42
HAU-net	Vgg16	**76.92**	**88.71**	**0.810**	**6.15**	**99.76**

2) RESULTS ON THE OPEN ENVIRONMENT DATASET

The test results for the open environment are shown in Table 4, and the results show that the seven models tested in this paper showed significant degradation in several evaluation metrics due to the interference of different backgrounds in the open environment. HPAM can weight the targets for the desired segmentation in space and channel, which mitigates the interference of backgrounds, so HAU-net still achieved the best results in several metrics on the open environment dataset results. It achieves 69.31 on MIoU, which is significantly better than other models. Also on Recall, HAU-net achieves 80.87, which is the highest among all models. In terms of the Dice coefficient, HAU-net also outperforms other models (0.806). In terms of HD, HAU-net is 6.22, which is better than less than the other models, so HAU-net still achieves the best results in several metrics on the open environment dataset.

TABLE 4. Performance of different models under open environment datasets.

Models	Backbone	MIoU	Recall%	Dice	HD	Accuracy%
U-net	Vgg16	69.27	79.77	0.796	6.29	97.72
U-net	Resnet50	68.69	80.80	0.790	6.54	97.74
U-net++		64.3	79.3	0.789	7.88	97.58
DeeplabV3		67.02	79.06	0.757	7.35	97.71
Segnet		55.45	54.82	0.548	7.89	97.56
FRCnet		60.59	71.76	0.655	7.57	97.59
HAU-net	Vgg16	**69.31**	**80.87**	**0.806**	**6.22**	**97.75**

3) RESULTS ON THE NON-OPEN ENVIRONMENT DATASET

The test results for the non-open environment dataset are shown in Table 5, and the results indicate that the HAU-net model still shows the best performance in all these metrics. Specifically, it achieves 76.05 on MIoU, which is significantly better than the other models. Also in Recall, HAU-net reached 88.14, which is the highest among all models. In terms of the Dice coefficient, HAU-net also outperformed the other models (0.847) and was second only to the Resnet50-based U-net model. Finally, HAU-net achieved 97.79 in Accuracy and 4.27 in HD, that are both the highest among all models.

Combining the performance of all three datasets, HAU-net shows excellent performance. It indicates that in the tongue crack segmentation scenario facing background interference and poor foreground hind scene separation, HAU-net improved with HPAM can overcome these problems and maintain good generalization performance.

TABLE 5. Performance of different models under non-open environment datasets.

Models	Backbone	MIoU	Recall%	Dice	HD	Accuracy%
U-net	Vgg16	74.56	86.02	0.843	4.29	99.69
U-net	Resnet50	75.43	84.97	**0.855**	4.27	99.05
U-net++		74.86	86.3	0.825	5.81	99.72
DeeplabV3		69.7	76.94	0.722	5.47	98.46
Segnet		68.09	75.1	0.744	5.92	99.31
FRCnet		68.88	77.96	0.739	5.41	99.2
HAU-net	Vgg16	**76.05**	**88.14**	0.847	**4.27**	**99.79**

4) IMAGES OF MODEL PREDICTION RESULTS

Figure 9 illustrates the effectiveness of the proposed model compared to various other models in the context of tongue cleft extraction. To ensure a fair comparison, we carefully selected test set images within each category for evaluating segmentation outcomes. Specifically, subsets 'a' and 'b' were chosen from non-open environment datasets, subsets 'c' and 'd' from open environment datasets, and subset 'e' from external data sources [27].

As depicted in the figure, the SegNet model tends to produce more cluttered artifacts within the original image during prediction. U-netpp exhibits omissions in images 'a' and 'd', while FRCnet displays omissions in image 'd' as well. Notably, the SegNet model consistently generates more cluttered islands in the original image during prediction.

In Figure 8, column 'c,' it's evident that each model exhibits varying degrees of omissions, primarily due to the intricate and challenging nature of tongue fissures in image 'c.' However, HAU-net demonstrates superior generalization capabilities, with its segmentation results in column 'c' closely aligning with the original labels.

In the case of image 'e,' Deeplabv3 and HAU-net exhibit superior performance, while other models introduce additional noise in their predictions. The performance in image 'e' serves as an indicator of the models' generalization capabilities.

The combined results from subsets 'a' to 'e' demonstrate that the model consistently excels in tongue cleft extraction, reflecting its strong overall performance.

F. INFERENCE TIME AND PARAMETER COUNT RESULTS

In addition to accuracy, the size and detection speed of the model are also of significant importance, especially for the task of tongue fissure extraction. To assess whether

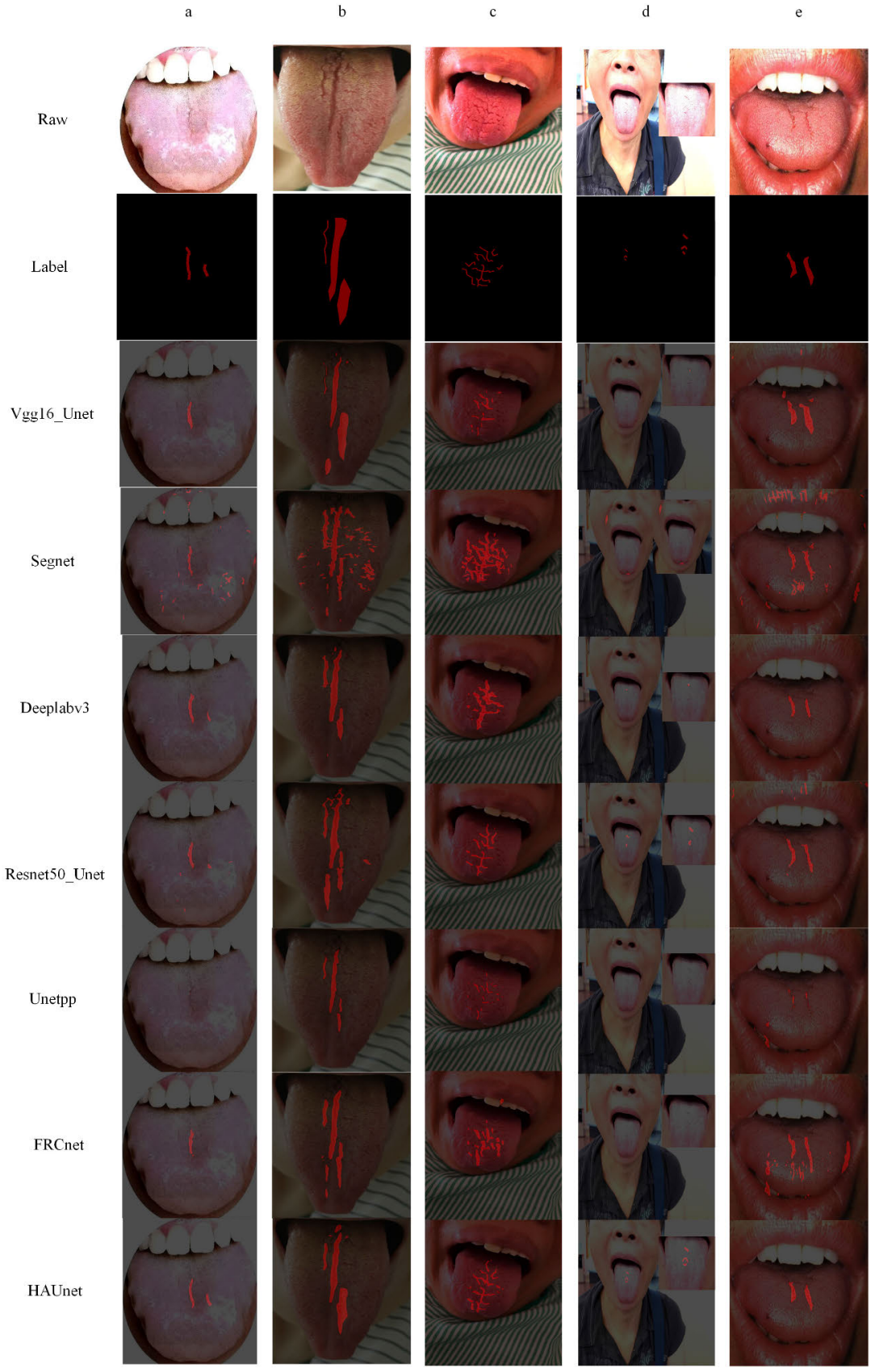

FIGURE 9. Images of different models prediction results.

the algorithm's detection speed can achieve real-time detection, we conducted tests on the average detection speeds of different models on the test set, and the test results are detailed in Table 6. From the perspective of inference time, HAU-net, U-net_resnet50, U-net_vgg16, DeeplabV3, and Segnet demonstrate similar inference speeds. However, in terms of model parameter size, both DeeplabV3 and Segnet have higher parameter counts compared to HAU-net, with U-net_resnet50 having an even higher parameter count of 43.93M. In contrast, FRCnet stands out with the lightest inference speed and parameter count, but this comes at the cost of a significant loss in accuracy, making it less suitable for the task of tongue fissure extraction.

The proposed HAU-net in this paper maintains stable inference speed while moderately increasing the parameter count and exhibits superior accuracy. In practical applications of fissure extraction, HAU-net efficiently performs accurate work.

TABLE 6. Inference time and parameter count results.

Models	Inference Time(ms)	Params
U-net_vgg16	15.76 ms	24.89M
U-net_resnet50	15.63 ms	43.93M
U-net++	16.02 ms	9.6M
DeeplabV3	15.66 ms	39.63M
Segnet	16.84 ms	29.94M
FRCnet	3.68 ms	**0.78M**
HAU-net	15.63 ms	24.93M

G. ABLATION EXPERIMENTS

To verify whether the addition of HPAM to U-net is due to performance optimization of a single module of either or a combination of multiple modules, this paper validates the effectiveness of the role of multiple hybrid attention mechanisms in series and in parallel. The text designs several comparative architectures:

i. Add CAM to the U-net decoder without adding SENet and SAM, which is noted as CAM in Table 6.

ii. Add SAM to the U-net decoder without adding CAM and SENet, and record it as SAM in Table 6.

iii. Add SENet to the U-net decoder without adding CAM and SAM, and record as SE in Table 6.

iv. Add SAM and CAM to the U-net decoder without adding SENet, and record as SAM in Table 6.

v. Add SENet and CAM to the U-net decoder without adding SAM, and record as SAM_CAM in Table 6.

vi. Add SENet and SAM to the U-net decoder without adding CAM, which is recorded as SE_CAM in Table 5.

vii. Add SENet, SAM and CAM in U-net decoder in tandem, noted as SE_SA_CA(S) in Table 6.

viii. Add SENet, SAM and CAM in parallel to the U-net decoder, which is recorded as SE_SA_CA(P) in Table 6.

ix. Add SENet, SAM, and CAM in parallel to the encoder of U-net, which is noted as SE_SA_CA(P)_EN in Table 5.

The structure of the ablation experiment is shown in Table 6:

TABLE 7. Results of ablation experiments.

Models	MIoU	Recall%	Dice	Accuracy%
CAM	75.56	0.807	87.25	99.74%
SAM	69.24	0.748	79.24	99.66
SE	76.43	0.809	87.63	99.76
SAM_CAM	75.70	0.800	87.72	99.74
SE_SAM	76.53	**0.814**	87.73	99.76
SE_CAM	75.4	0.803	85.67	99.75
SE_SA_CA(S)	75.35	0.808	83.42	99.73
SE_SA_CA(P)_EN	74.6	0.801	**82.82**	**99.71**
SE_SA_CA(P)	**76.92**	0.810	**88.71**	**99.76**

From the ablation experiments presented in Table 5, the following conclusions can be drawn:

1. When using individual attention mechanisms alone, the SENet performs the best, achieving a MIoU of 76.43% on the test set. This is likely because SENet learns channel dependencies through the introduction of the SE block, which adaptively adjusts the weights of each channel in the feature map. This enables the model to focus on the most relevant features and enhances its generalization ability.
2. When combining attention mechanisms, the combination of SE and SAM (SE+SAM) yields the highest Dice metric of 0.814. This is possibly due to SAM learning pixel relationships to adjust the weights of each pixel in the tongue crack image, complementing the attention mechanism of the SE channel. The combination of these two mechanisms produces synergistic effects, resulting in improved performance.
3. When using all three attention networks simultaneously, the MIoU of the SE+SAM+CAM model with parallel attention mechanisms (P) across three branches surpasses the MIoU of the three attention mechanisms in series (S), reaching 76.92%. This is the highest performance among all the compared models. The parallel structure allows for better performance as each attention mechanism addresses different aspects. SENet learns channel dependencies, SAM focuses on spatial attention, and CAM emphasizes channel attention. In contrast, using these three mechanisms in series may lead to incomplete information or interference, reducing network performance. Thus, the ablation experimental results indicate that the parallel structure is a preferable choice for achieving better results in the tongue crack extraction task.

Furthermore, to visually explore the impact of HPAM on the model's crack extraction ability, weights from the last layer of each ablation model were extracted using the Grad-CAM [22] technique to generate heatmaps. As shown in Figure 10, the tongue crack image in the open environment

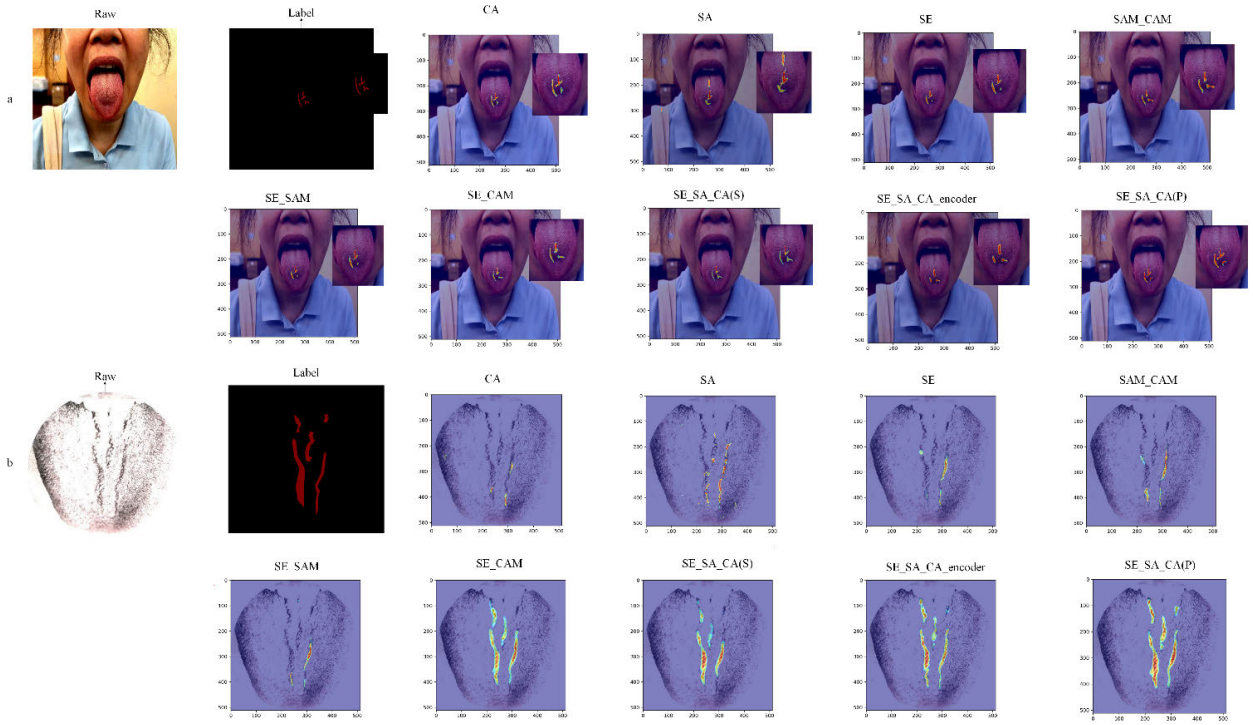

FIGURE 10. Heat map of attention for ablation experiments.

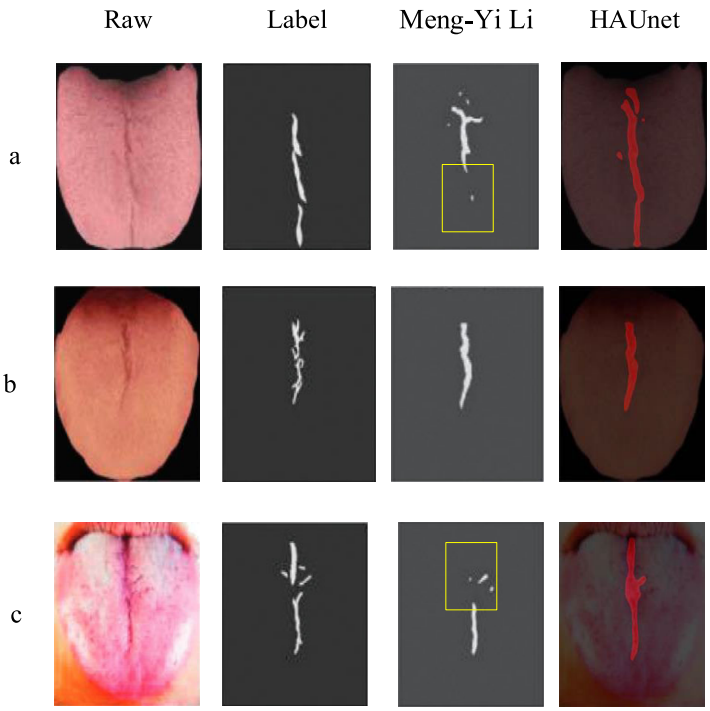

FIGURE 11. Tongue crack extraction performance comparison chart.

(a) and non-open environment (b) are presented. In the SE+SA+CA(P) model with HPAM added, the weight of the tongue crack region is significantly higher compared to other comparison models. Heatmaps of the SE and SE_SAM segmentations reveal lower weights (lighter color) in the marked yellow box region, indicating that the model fails to extract the tongue crack correctly in this region, resulting in reduced accuracy and missed detections. In contrast, in the SE+SA+CA(P) model, the weight at the edge of the tongue crack image increases, represented by a darker red color.

This indicates successful extraction of tongue cracks, even in challenging anterior and posterior views.

IV. DISCUSSION

The main goal of this study is to design and test a U-net model embedded with HPAM to improve the accuracy of tongue crack extraction. Our experimental results strongly demonstrate the superiority of this structure in handling the tongue cracking task.

In comparison experiments, our model HAU-net significantly outperforms the original U-net, U-netpp, Deeplabv3, Segnet, and FRCnet models in key performance metrics such as recall, precision, MIoU, and Dice values. These results validate the superior performance of our model in identifying and segmenting tongue cracks, which may be attributed to its ability to effectively utilize the multi-track parallel attention mechanism to extract and exploit richer and more complex features. To further validate the generalization performance of the HAU-net model, we used the graph of Meng-Yi Li's paper and the experimental results [11] as a comparison, as shown in Figure 11, where the method proposed by Meng-Yi Li has significant missed detection in both a and c, marked by yellow boxes in the graph. Although we use Meng-Yi Li's image annotation, the crack extraction results of the method in this paper, in Fig. 11c, fit the original image more closely and are even more accurate than the annotated image. In comparison, the HAU-net method has a significant improvement in the tongue crack extraction task.

For practical applications and future research, this improved model may have far-reaching implications. First, due to its excellent performance in tongue fissure extraction, this model has the potential to further improve the accuracy and efficiency of objectified diagnosis in TCM tongue diagnosis. Second, by showing that more complex attentional mechanisms can effectively improve model performance, our study may encourage more researchers to incorporate and explore this novel structure in their models.

Although our model achieves good results in extracting tongue cracks, we believe there is still room for further improvement. First, we hope to further improve the performance of the model by optimizing and tuning the multi-track parallel attention mechanism. Second, we also hope to apply this model to other image segmentation tasks in the future to explore its possibilities in a wider range of domains. We also plan to collect and build a larger dataset of tongue cracks to fully validate the robustness and generalization ability of our model.

V. CONCLUSION AND FUTURE WORK

(1) To develop a suitable model for tongue crack extraction, this study proposes the HAU-net model by adding the HPAM module to the decoder network structure of the U-net. Compared with the other seven tested models, HAU-net all showed different degrees of improvement. HAU-net achieved the highest MIoU of 76.92, recall of 88.71%, accuracy of 99.76%, and Dice of 0.810 on the overall data set. the MIoU improved by 1.73%; compared with the original U-net model, the MIoU and DICE were significantly improved, and the number of model parameters and inference rate did not change significantly.

(2) The results of the ablation experiments show that the enhancement of the model by adding HPAM to the U-net decoder section is the most obvious. And the HPAM structure is more effective in parallel than in series.

(3) This study not only provides an effective method for the automatic extraction of cracked tongue, but also contributes to the automation and accuracy of tongue diagnosis. This may help to improve and optimize the process of TCM diagnosis, especially for tongue diagnosis, which is an important diagnostic component.

(4) In our future work, we will explore and assess the application of various advanced attention mechanisms to enhance model performance. This may include multi-head attention, cross-modal attention, and more. We will delve into these techniques and endeavor to integrate them into our model to improve its performance in tongue fissure extraction tasks. Additionally, we will continue to seek out additional tongue fissure datasets and explore various data augmentation and enhancement techniques to further enhance the model's robustness and generalization capabilities.

REFERENCES

[1] M. Sharma and V. K. Sharma, "Recurrent facial palsy and fissured tongue," *Eur. J. Internal Med.*, vol. 89, pp. 104–105, Jul. 2021, doi: 10.1016/j.ejim.2021.03.007.

[2] R. Sudarshan, G. S. Vijayabala, Y. Samata, and A. Ravikiran, "Newer classification system for fissured tongue: An epidemiological approach," *J. Tropical Med.*, vol. 2015, Sep. 2015, Art. no. e262079, doi: 10.1155/2015/262079.

[3] C. Casu, G. Mosaico, V. Natoli, A. Scarano, F. Lorusso, and F. Inchingolo, "Microbiota of the tongue and systemic connections: The examination of the tongue as an integrated approach in oral medicine," *Hygiene*, vol. 1, no. 2, pp. 56–68, Jul. 2021, doi: 10.3390/hygiene1020006.

[4] Q. Li, Y. Wang, H. Liu, Z. Sun, and Z. Liu, "Tongue fissure extraction and classification using hyperspectral imaging technology," *Appl. Opt.*, vol. 49, no. 11, pp. 2006–2013, 2010, doi: 10.1364/AO.49.002006.

[5] X. Wang, J. Liu, C. Wu, J. Liu, Q. Li, Y. Chen, X. Wang, X. Chen, X. Pang, B. Chang, J. Lin, S. Zhao, Z. Li, Q. Deng, Y. Lu, D. Zhao, and J. Chen, "Artificial intelligence in tongue diagnosis: Using deep convolutional neural network for recognizing unhealthy tongue with tooth-mark," *Comput. Struct. Biotechnol. J.*, vol. 18, pp. 973–980, Jan. 2020, doi: 10.1016/j.csbj.2020.04.002.

[6] W. Tang, Y. Gao, L. Liu, T. Xia, L. He, S. Zhang, J. Guo, W. Li, and Q. Xu, "An automatic recognition of Tooth- marked tongue based on tongue region detection and tongue landmark detection via deep learning," *IEEE Access*, vol. 8, pp. 153470–153478, 2020. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9171233

[7] R. Kanawong, T. Obafemi-Ajayi, T. Ma, D. Xu, S. Li, and Y. Duan, "Automated tongue feature extraction for Zheng classification in traditional Chinese medicine," *Evidence-Based Complementary Alternative Med.*, vol. 2012, pp. 1–14, Jan. 2012, doi: 10.1155/2012/912852.

[8] Z. Huang, J. Miao, H. Song, S. Yang, Y. Zhong, Q. Xu, Y. Tan, C. Wen, and J. Guo, "A novel tongue segmentation method based on improved U-Net," *Neurocomputing*, vol. 500, pp. 73–89, Aug. 2022.

[9] Q. Ruan, Q. Wu, J. Yao, Y. Wang, H.-W. Tseng, and Z. Zhang, "An efficient tongue segmentation model based on U-Net framework," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 35, no. 16, Dec. 2021, Art. no. 2154035.

[10] H. Song, Z. Huang, L. Feng, Y. Zhong, C. Wen, and J. Guo, "RAFF-Net: An improved tongue segmentation algorithm based on residual attention network and multiscale feature fusion," *Digit. Health*, vol. 8, Jan. 2022, Art. no. 205520762211363.

[11] L. Zhu, G. Xin, X. Wang, C. Ding, H. Liang, and Q. Chen, "A fast tongue detection and location algorithm in natural environment," *Comput., Mater. Continua*, vol. 73, no. 3, pp. 4727–4742, 2022.
[12] H. Weng, L. Li, H. Lei, Z. Luo, C. Li, and S. Li, "A weakly supervised tooth-mark and crack detection method in tongue image," *Concurrency Comput., Pract. Exp.*, vol. 33, no. 16, Aug. 2021, Art. no. e6262.
[13] Y. Xue, X. Li, Q. Cui, L. Wang, and P. Wu, "Cracked tongue recognition based on deep features and multiple-instance SVM," in *Advances in Multimedia Information Processing—PCM 2018*. Cham, Switzerland: Springer, 2018, pp. 642–652. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-030-00767-6_59
[14] J. Yan, J. Cai, Z. Xu, R. Guo, W. Zhou, H. Yan, Z. Xu, and Y. Wang, "Tongue crack recognition using segmentation based deep learning," *Sci. Rep.*, vol. 13, no. 1, p. 511, Jan. 2023.
[15] M.-Y. Li, D.-J. Zhu, W. Xu, Y.-J. Lin, K.-L. Yung, and A. W. H. Ip, "Application of U-Net with global convolution network module in computer-aided tongue diagnosis," *J. Healthcare Eng.*, vol. 2021, pp. 1–15, Nov. 2021, doi: 10.1155/2021/5853128.
[16] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," 2023, *arXiv:2304.02643*.
[17] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16 × 16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
[18] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*. Munich, Germany: Springer, Oct. 2015, pp. 234–241.
[19] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020, doi: 10.1109/TPAMI.2019.2913372.
[20] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19. [Online]. Available: https://openaccess.thecvf.com/content_ECCV_2018/html/Sanghyun_Woo_Convolutional_Block_Attention_ECCV_2018_paper.html
[21] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," 2016, *arXiv:1606.04797*.
[22] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y.-W. Chen, and J. Wu, "UNet 3+: A full-scale connected UNet for medical image segmentation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Barcelona, Spain, May 2020, pp. 1055–1059. [Online]. Available: https://ieeexplore.ieee.org/document/9053405/
[23] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.
[24] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder–decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017, doi: 10.1109/TPAMI.2016.2644615.
[25] L. Shi, Y. Wang, Z. Li, and W. Qiumiao, "FRCNet: Feature refining and context-guided network for efficient polyp segmentation," *Frontiers Bioeng. Biotechnol.*, vol. 10, Jun. 2022, Art. no. 799541, doi: 10.3389/fbioe.2022.799541.
[26] 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 *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.
[27] T. B. Song, "Practical traditional Chinese medicine tongue diagnosis color atlas," Anhui Sci. Technol. Press, Hefei, China, Tech. Rep., 1994.

ZIHAO ZHANG received the B.S. degree in network engineering from Polytechnic Normal University, in 2020. He is currently pursuing the M.S. degree in computer science with the Zhongkai University of Agricultural Engineering. His research interests include new technologies and methods in Chinese medicine research and artificial intelligence.

JIANHUA ZHENG received the M.S. degree in machine design and theory and the Ph.D. degree in mechatronic engineering from the South China University of Technology, Guangzhou, China, in 2002 and 2010, respectively.

From 2011 to 2013, he did his postdoctoral research in computer science and technology. He is currently an Associate Professor with the Zhongkai University of Agricultural Engineering. He mainly focuses on artificial intelligence application in different areas, such as medical and agriculture.

RUOLIN ZHAO received the B.S. degree in electronic information science and technology from Henan University, in 2021. She is currently pursuing the M.S. degree in agricultural engineering and information technology with the Zhongkai University of Agricultural Engineering. Her research interests include image processing, machine vision, and artificial intelligence.

SHUANGYIN LIU received the Ph.D. degree from the College of Information and Electrical Engineering, China Agricultural University, in 2014. He is currently a Professor with the College of Information Science and Technology, Zhongkai University of Agriculture and Engineering. His current research interests include the areas of intelligent information systems for agriculture, artificial intelligence, bigdata, and computational intelligence.

ZHENGJIE LIU received the bachelor's and master's degrees from the Jiangxi College of Traditional Chinese Medicine in 1994 and 1999, respectively, and the Ph.D. degree in traditional Chinese medicine from the Guangzhou University of Traditional Chinese Medicine, in 2002.

He is currently the Director of the Department of Endocrinology, Guangdong Provincial Hospital of Traditional Chinese Medicine, the Chief Physician, and a Postgraduate Tutor. He specializes in the integrated traditional Chinese and western medicine treatment of diabetes and other endocrine and metabolic diseases.

JINHE WANG received the B.S. degree in medicine, in 2021. He is currently pursuing the degree in the clinical specialty of integrated traditional and western medicine with the Graduate School, Chinese Academy of Traditional Chinese Medicine. His research interests include integrated traditional and western medicine skin and venereology, and he is engaged in related clinical and basic research.