scientific reports

Check for updates

OPEN

An improved lightweight tongue segmentation model with self-attention parallel network and progressive upsampling

Xuan Wang[1], Yifang Cao[1], Yijia Chen[2], Huixia Li[3]✉, Aiqing Han[1]✉ & Yan Tang[1]✉

Tongue diagnosis is a crucial component of the Four Diagnostic Methods in Traditional Chinese Medicine (TCM), which include observing, listening and smelling, inquiring, and palpation. Tongue image segmentation holds great significance in advancing the intelligentization of tongue diagnosis research. This paper introduces an improved model called Parallel Attention and Progressive Upsampling for Tongue Segmentation (PAPU_TonSeg), based on the Segformer architecture, to address the issues of inaccurate and blurred tongue edge segmentation in tongue semantic segmentation. The model incorporates three key enhancements: (1) the adoption of a Self-Attention Parallel Network that integrates the self-attention mechanism and residual modules to achieve simultaneous extraction of local and global features; (2) the integration of the Efficient Channel Attention(ECA) mechanism into the Mix-FFN component to enhance feature extraction efficiency; and (3) the utilization of Multi-dimensional Feature Progressive Upsampling to mitigate precision loss during the upsampling process. Evaluation results on the BioHit public dataset demonstrate that, compared to the original Segformer, PAPU_TonSeg achieves improvements of 2.42% in Mean Pixel Accuracy (MPA), 0.78% in Mean Intersection over Union (MIoU), and 2.02% in the Dice coefficient, while boasting a lower parameter count and computational complexity. On another dataset, PAPU_TonSeg outperforms Segformer with an MPA increase of 0.64%, an MIoU increase of 0.33%, and a Dice coefficient increase of 0.4%. The improved model not only has fewer parameters but also exhibits a notably lower computational complexity compared to classical models. The PAPU_TonSeg model accurately segments tongue body details, such as tooth marks, and distributes attention more evenly, capturing both global and local features. These findings position PAPU_TonSeg as a valuable tool for clinical diagnosis and research in TCM tongue diagnosis.

Keywords Tongue semantic segmentation, Parallel network, Multi-level feature aggregation, The ECA attention mechanism, Segformer

Tongue diagnosis is a significant part of the four diagnostic methods in traditional Chinese medicine (TCM), which include inspection, listening and smelling, inquiry, and feeling the pulse. By combining the tongue diagnosis with pulse diagnosis, inquiry, and other diagnostic methods, TCM doctors can correctly differentiate the patterns, and prescribe appropriate treatments for patients. In TCM, the tongue is connected to the internal organs (zang-organs and fu-organs) through the meridians and collaterals; thus, the conditions of the organs, qi, blood, and bodily fluids, as well as the degree and progression of disease, are manifested in the tongue. Tongue diagnosis can help TCM doctors differentiate the patterns, such as distinguishing between cold patterns and heat patterns or deficient patterns and excess patterns[1].

Tongue diagnosis of TCM can observe the patients' condition from many aspects, namely the motility of the tongue, tongue coating, and tongue color, which provide insights into the internal state of the body and organs. Firstly, the motility of the tongue reflects the severity of diseases. For example, a healthy tongue that indicates vigorous organ function is characterized by flexibility and the ability to extend and retract appropriately while a pathological tongue may manifest as flaccidity, rigidity, deviation, and so on. Secondly, the tongue coating is a

[1]School of Management, Beijing University of Chinese Medicine, Beijing, China. [2]College of Humanities, Beijing University of Chinese Medicine, Beijing, China. [3]Department of Spleen and Gastroenterology, The Third Affiliated Hospital of Beijing University of Chinese Medicine, Beijing, China. ✉email: lihuixiadeemail@163.com; aqhan@hotmail.com; tangyan97_1017@sina.com

thin, moist layer on the surface of the tongue, consisting of a mixture of epithelial cells, saliva, microorganisms, and food debris[2]. A healthy tongue coating is a thin white layer that is evenly distributed over the surface of the tongue. It is produced by the stomach qi. The patient's thick, yellow coating is often indicative of excess fire and stagnation in the spleen and stomach. Thirdly, different tongue color reflects different body condition. A light red tongue is generally considered healthy; while redness at the tip of the tongue often signifies upward flaming of heart fire, leading to patterns such as vexation, insomnia, and ulcers on the mouth and tongue. The moist pale purple or bluish-purple tongue usually indicates yin-cold becomes excessive inside and stagnation in the blood vessels. Therefore, it can be concluded that the state of disease is closely related to the tongue's motility, coating, and color[3].

Integrating TCM theories with advanced technologies such as information technology and Artificial Intelligence to achieve the automation and intelligence of the four diagnostic methods of TCM, thereby promoting the innovative development of TCM[4]. Currently, the subjective interpretation of TCM doctors, along with environmental factors such as lighting and surroundings, can impact the neutrality and precision of tongue diagnosis, which is one of the key components of the Four Diagnostic Methods. To overcome the limitation of traditional tongue diagnosis and advance towards intelligent tongue diagnosis, it is necessary to construct a Deep Learning-based tongue image diagnosis model, which consists of tongue semantic segmentation and tongue image classification. The tongue semantic segmentation requires to elimination of non-tongue areas such as the face and lips from the images, leaving only the tongue body to enhance the accuracy of tongue image classification. Therefore, a high-precision tongue semantic segmentation is an essential component of intelligent tongue diagnosis.

The conventional tongue image segmentation methods mainly utilize statistical machine learning and image processing technologies to achieve segmentation. Bo et al. proposed an original technique based on the combination of a bi-elliptical deformable template and an active contour model, which captures gross shape features through an energy function to segment tongue images[5]. Shi et al. introduced color space information to control the evolution of curves. They combined the geometric snake model with the parameterized GVFSnake model and proposed a new method for automatic tongue image segmentation that can achieve automatic segmentation of the tongue body[6]. Shi et al. digitized tongue body images and proposed a double geo-vector flow method for detecting tongue edges and segmenting the tongue region in the images, achieving a certain effect in tongue body segmentation[7]. Conventional tongue image segmentation is marked by high computational complexity and the need for geometric or mathematical modeling for image analysis. Some methods even require manual extraction of the tongue's color, texture, and shape. Moreover, many models are easily influenced by the changes in lighting and are unable to distinguish the tongue from the lips. Therefore, there remains a great challenge to segment tongue images by using existing technique methods[8].

In the field of image processing, numerous innovative approaches have emerged to address various challenges, which can be broadly categorized as follows: Leverages SAM to segment ambiguous objects by exploiting its potential in uncertain regions[9]; Enhances SAM with probabilistic prompting for efficient segmentation of ambiguous medical images[10]; Proposes modality-prompted heterogeneous graph learning for omni-modal biomedical representation[11]; Integrates global correlation networks and discriminative embedding for few-shot segmentation[12]; Unsupervised anomaly segmentation: Detects brain lesions via dual semantic-manifold reconstruction[13]; Synthesizes pseudo-healthy images via GANs to refine lesion segmentation[14].

With the development of AI Deep Learning, tongue semantic segmentation which is based on Deep Learning is increasingly gaining attention. As proposed by Long et al., the FCN (Fully Convolutional Network) introduced deep learning algorithms into the field of image segmentation[15]. Xu et al. employed the FCN for tongue semantic segmentation, thereby realizing a substantial improvement in accuracy in contrast to conventional methodologies[16]. Zhou et al. proposed an end-to-end deep neural network model, which incorporates a feature pyramid network of residual blocks designed for the extraction of multi-scale tongue features. By leveraging the extracted feature maps, the model is capable of pinpointing candidate tongue regions and subsequently executing precise localization and segmentation of the tongue, thereby yielding impressive segmentation results[17]. Based on FCN, Ronneberger et al. presented the U-Net Neural Network. It can extract features at different resolutions through its unique U-shaped network structure and fuses these features with upsampled features in the decoding part[18]. U-Net is one of the most widely applied approaches in deep learning for image segmentation, which exhibits good performance in segmenting biomedical images[19]. Xu Q et al. suggested a multi-task joint learning method that combines a U-Net segmentation network model with a discriminative filter learning model to achieve tongue image segmentation and classification tasks[20]. To better extract high-level semantic features, Li M Y et al. introduced the Global Convolution Network Module into the encoder part of U-Net. An improved U-Net network was designed for the semantic segmentation of fissured tongues, which segments tongue fissures effectively[21].

In recent years, the rapid development of computer vision in Deep Learning has led to many semantic segmentation models and their variants. Those models have achieved very good results in the field of medical image segmentation[22]. Building upon the encoding phase's edge information and by integrating attention mechanisms, Liu et al. proposed a multi-layer edge attention network to comprehensively extract features from medical images. This network has demonstrated superior segmentation performance on Tongue Image Segmentation Dataset[23]. Qiu et al. conducted segmentation and classification of the tongue body on mobile devices. They implemented a high-precision segmentation technique that leveraged data augmentation and moment invariants on the dataset. They then integrated an attention mechanism after their segmentation, to facilitate lightweight tongue image classification, which yielded promising results[24]. The Deeplab series of neural networks introduced Atrous Convolution, enabling the expansion of the receptive field without a corresponding increase in the number of parameters[25]. Chen et al. put forward the DeepLabv3 model by improving the pyramid-shaped Atrous Pooling, cascading dilated convolutions, and making extensive use of

batch normalization[26]. Zhang et al. have enhanced the decoder of the DeepLabV3 network, thereby significantly mitigating the ambiguity in tongue image segmentation. Additionally, a post-processing module has been appended to augment the segmentation precision in the regions proximate to the tongue edges as well as those remote from the tongue edges[27].

The above studies have all achieved considerable outcomes of tongue segmentation. However, the current tongue image segmentation models still suffer from the following deficiencies: ① the network models which are mainly based on convolutional neural networks fail to extract the global context semantic information of pictures effectively. In realistic segmentation situations, there will be occurrences such as incorrect judgments or the lack of certain parts of the tongue image. ② The semantic segmentation models taking the Transformer as the nucleus must rely on the Positional Embedding architecture to supply position information so that the limitations of the Self-Attention mechanism in processing the position information of different Patches can be overcome. However, when the resolution of the input image changes, the Positional Embedding with a fixed resolution is unable to accurately transmit the position information of Patches, which in turn causes a loss of accuracy. Meanwhile, the huge number of parameters in the Transformer model requires a large amount of data for training, and insufficient data will affect its performance optimization. ③ Most of the mainstream segmentation models are adopted to an encoding–decoding structure. In most of these models, the feature maps in the decoding process must be subjected to 16-times bilinear upsampling to recover the original size. This process is apt to cause a loss of edge accuracy, and ultimately leads to blurry edges in the segmented pictures, making it difficult to meet the application requirements of traditional Chinese medicine tongue diagnosis.

To address the above-mentioned problems, we propose PAPU_TonSeg (Parallel Attention and Progressive Upsampling for Tongue Segmentation), a high-precision, lightweight tongue image segmentation network designed for tongue image segmentation. This network incorporates a parallel structure that leverages self-attention mechanisms along with progressive upsampling techniques. To achieve this improvement, we conducted an in-depth study on the network structures of semantic segmentation models and tested various models. Eventually, we chose to enhance the Segformer model, specifically for tongue segmentation in practical scenarios. The innovation and highlights of this paper mainly include the following three points.

(1) Incorporate a parallel network module into the Segformer. While obtaining the global semantic features, fuse the high-dimensional detail features as well.
(2) To make the Transformer structure extract semantic and feature information more effectively, the Transformer structure is combined with the convolutional model, and the ECA attention mechanism is added, aiming to achieve more refined feature weight allocation and optimize the representational ability of the model.
(3) To reduce the loss of accuracy during the upsampling process, a structure that combines progressive upsampling with a convolutional network model is adopted. This structure is designed to alleviate the loss of accuracy during the upsampling process.

This paper consists of four sections. The first section is a brief introduction. The second section presents the introduction of relevant methods and model improvements. The third section focuses on the experiments and experimental results. The last section is dedicated to the discussion and summary.

Methods and model improvements

This section includes two parts. The first part briefly introduces the relevant theories of the Segformer. The second part presents the improvement ideas of the Segformer and the overall structure of the improved model.

Segformer model theory

In 2017, the Google team brought in the revolutionary Transformer model within the realm of artificial intelligence translation. Transformer constructed a novel network architecture by integrating the core components such as the innovative Self-Attention mechanism, Multi-Head Attention strategy, and Positional Encoding[28]. The innovation on the structural aspect empowers the Transformer to manifest outstanding performance in natural language processing (NLP) tasks, which swiftly emerging as the leading technology within this domain. The most prominent feature of the Transformer model lies in its abandonment of legacy CNN and RNN, with the self-attention mechanism being taken as the core of the model. While the Transformer structure has achieved considerable achievements in the field of natural language processing, it has also obtained considerable results in computer vision tasks. Kolesnikov et al. advanced the Vision Transformer (ViT) model predicated on the Transformer. It has obtained outstanding achievements in the realm of image classification, which demonstrate the Transformer's application value in the realm of Computer Vision. From then on, models based on the Transformer structure have continuously emerged in the fields of image classification, image detection, and image segmentation[29].

In 2021, Xie et al. proposed a Segformer model based on the Transformer. They improved the encoder and decoder of the Transformer through advanced ideas, namely the elimination of positional encoding, the employment of a multi-level Transformer encoder, and the simplification of an All-MLP decoder[30]. While simplifying the model structure, the model still achieved the best performance on segmentation tasks in various datasets at that time. In addition, compared with the conventional Transformer model, the Segformer can obtain satisfactory training effects with a relatively small quantity of data. The schematic of the Segformer model's structure is presented in Fig. 1.

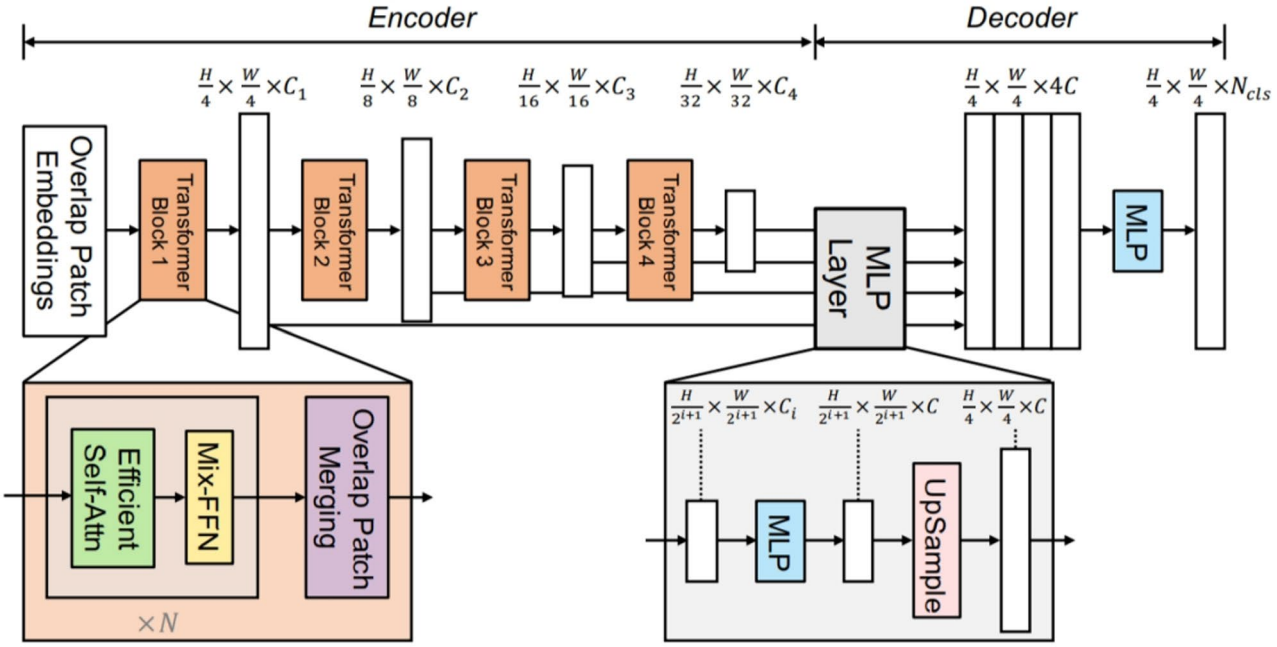

Fig. 1. Structure of Segformer.

Improved PAPU_TonSeg model

This study performs tongue segmentation under application situations based on the improved PAPU_TonSeg model. This section will present the improvement ideas of the model as well as the overall structure of the improved model.

Parallel network structure that combines the self-attention mechanism and residual module

A parallel network can run different network modules simultaneously and then combine their outputs. Its core concept is to make use of the computational power of multiple network modules to enhance efficiency and performance by handling data concurrently. Parallel networks can combine the advantages of two network structures. In the realm of convolutional neural networks, Szegedy et al. put forward the Inception-ResNet-v2 parallel network model in 2016, which combines the features of the ResNet residual connection mechanism that can prevent the disappearance of gradients and improve performance, and the characteristics of the Inception module that can obtain sparse features under the same dimension[31]. After the ViT model was proposed, relevant research efforts were made to integrate the merits of the Transformer and traditional Convolutional Neural Networks, thereby proposing a parallel network structure. The research team constituted by Microsoft and the University of Science and Technology of China put forward the Mobile-Former model. By paralleling the Mobilenet network with the Transformer network, this model shows considerable performance in the situation of simplified model parameters[32].

Compared with classic convolutional neural network layers, the Transformer structure possesses a broader scope of vision and is more capable of effectively capturing global characteristics. Through the employment of heatmap-based visualization techniques, the Segformer team delineated the effective perceptual scope of both Segformer and DeepLabv3 + models, noting that Segformer exhibited a more efficacious perceptual scope throughout the four phases of encoder subsampling. Even at the fourth layer of downsampling Segformer's perceptual field surpasses that of DeepLabv3 + , which is a key element contributing to the Segformer model's capacity to manage contextual information effectively. However, in some specific application scenarios of tongue image segmentation, the model must focus more on the fine details of the tongue body's edge. The reasons for that are the edge of the tongue body is similar in color to the oral cavity and palate, and the tooth marks on the tongue body edge are a critical component of tongue diagnosis. To enable the model to obtain both local and global features simultaneously, based on existing research, this paper proposes a parallel network structure that integrates self-attention mechanisms with residual blocks. This structure processes the input neural network layer's data tensors in both the self-attention mechanism and the residual module concurrently, and then integrates the outputs from the distinct networks and passes them into the following layer of the network structure.

Incorporating ECA attention mechanism into the mix-FFN

The attention mechanism significantly helps improve the performance of models. In the field of semantic segmentation, the attention mechanism can also help models achieve better evaluations. U-KAN, by using the KA activation function, significantly enhances the model's ability to capture edge details in image segmentation tasks, especially in scenarios with complex textures and small object segmentation[33]. In the article "Hierarchical

deep network with uncertainty-aware semi-supervised learning for vessel segmentation", the authors propose a hierarchical deep network that combines the attention mechanism and uncertainty-aware semi-supervised learning to address the challenges of vessel segmentation in medical images[34]. "Attention U-Net: Learning Where to Look for the Pancreas" introduces a spatial attention gate in the skip connections of U-Net to automatically focus on the target area (such as the pancreas) and suppress irrelevant background[35]. "SA-Net: A Scale-Attention Network for Medical Image Segmentation" combines multi-scale features and channel attention to solve the problem of large differences in target sizes in medical images[36].

The Squeeze-and-Excitation (SE) channel attention mechanism is a typical attentional framework, wherein it modulates the salience of channel-wise features through the acquisition of their underlying nonlinear interplays[37]. Yet, the presence of fully connected layers in the SE module significantly increases the model's parameter count and computational complexity. To overcome this problem, Wang et al. introduced the Efficient Channel Attention (ECA) mechanism. Differing from the SE channel attention mechanism that relies on fully connected layers, the ECA attention mechanism opts for a 1×1 convolutional layer to diminish the model's parameter volume. Moreover, this approach allows the model to maintain a lower computational complexity, thereby enhancing performance without a concomitant increase in computational cost. The structure of the ECA attention mechanism is presented in Fig. 2.

The Segformer's core structure incorporates the Mixed Feedforward Network (Mix-FFN), which is an essential component in the fusion and extraction of features. This component is entirely constructed from convolutional neural networks, which are responsible for feature mixing and extraction. To further enhance the network's capability to extract features, we have incorporated the ECA attention mechanism within the Mix-FFN framework. By harnessing the strengths of the ECA attention mechanism in tandem with the Segformer network's inherent self-attention mechanism, the crucial features can be more effectively distilled from the data. This enhancement not only elevates the model's performance metrics but also curtails the training timeframe and computational burden. As a result, the model's utility and efficacy are markedly enhanced.

Multi-level feature aggregation

Semantic Segmentation Models Typically Employ an Encoder-Decoder Architecture. The architecture prevalent in semantic segmentation models facilitates feature extraction through the encoder phase, followed by an upsampling procedure by the decoder to reconstruct the image to its pristine size for precise pixel-level categorization. However, in this architecture, the feature maps produced by the decoder are usually much smaller in size than the original input image, implying that the upsampling process must amplify the coarse feature maps to the original image's size. This implies that during the upsampling process, the model needs to restore the coarse feature maps to the same size as the original image, a process that is prone to losing detailed information, leading to a decrease in the precision of the predicted features.

To minimize the loss of detailed information during the process of upsampling, we leverage the Segformer encoder's output of feature maps at various scales and employ a Multi-Level Feature Aggregation (MLA) upsampling approach. This method involves the sequential upsampling of feature maps of different dimensions to the required size using both convolutional neural networks and bilinear interpolation, with each iteration doubling the magnification factor. The disparate features are then merged. Notably, the MLA technique is devoid of complex neural network computations, thus not leading to a sudden increase in the model's parameter count.

Improved model structure

The improved PAPU_TonSeg model's network architecture consists of four cascaded Transformer units, a lightweight Multilayer Perceptron (MLP) decoder, parallel network modules, ECA attention mechanisms, and a Multi-Level Feature Aggregation (MLA) upsampling approach. And incorporating parallel network modules, ECA, and MLA upsampling sequentially based on the foundational Segformer model. The network structure of the upgraded PAPU_TonSeg model is depicted in Fig. 3.

As depicted in Fig. 3, the PAPU_TonSeg model's architecture has been refined by transforming the initial four Transformer units into four Parallel Blocks. These Parallel Blocks consist of parallelly operating Resnet Layers and Transformer Blocks. Furthermore, we have incorporated the ECA attention mechanism into the

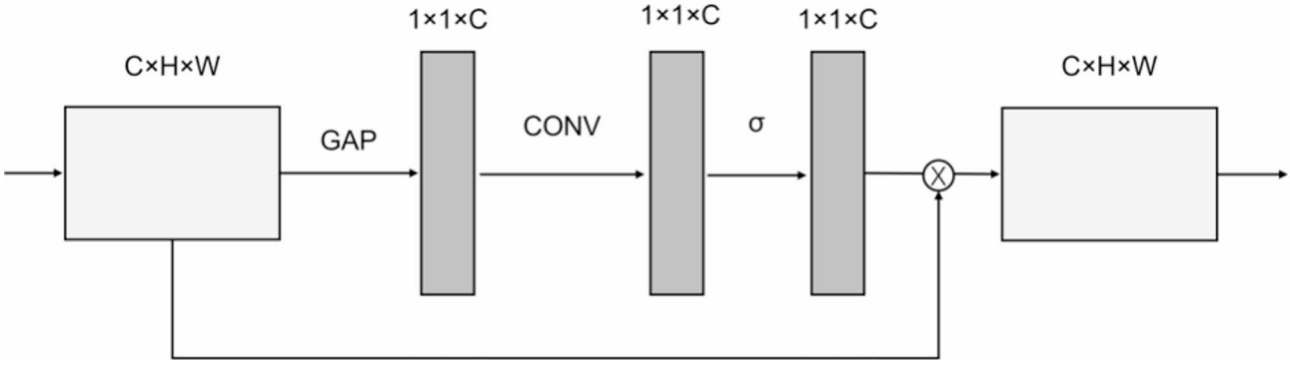

Fig. 2. Structure of the ECA attention mechanism.

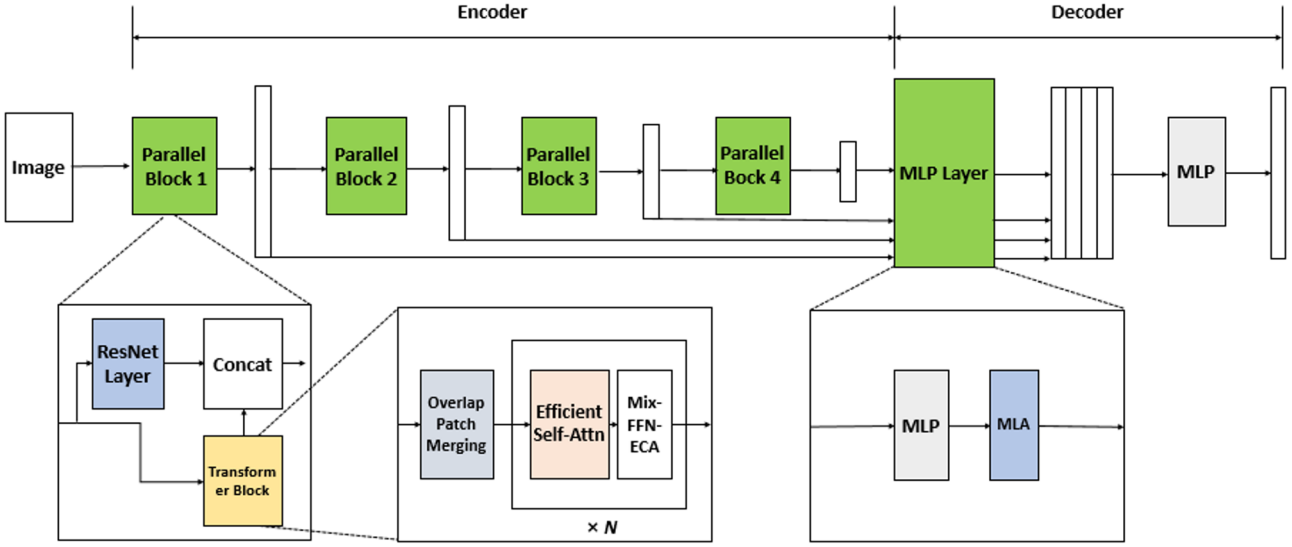

Fig. 3. Improved PAPU_TonSeg network architecture diagram.

Mix-FFN within the Transformer Block module. The Decoder section of the original model has been upgraded by replacing the MLP layer with a synergistic MLP and MLA setup, employing both convolutional neural networks and bilinear interpolation for a multi-tiered feature aggregation upsampling process. The process involves directly linking feature maps of varying dimensions from the encoder's middle layers to the decoder and equalizing the feature map dimensions. Then upsampling each tier of features to a quarter of the original image size through a feedforward neural network followed by bilinear interpolation. The features from the four layers are concatenated to achieve feature fusion, and another network is used to transform the channel numbers to match the quantity required for the semantic segmentation task to facilitate subsequent mask prediction.

Input: $X \in R^{B\times3\times H\times W}$

Output: $\mathbf{F}_3$ //at 1/4 input resolution

1.$\mathbf{F}_0$=ReLU(BatchNorm(Conv2D(X,k=7,s=2,p=3,out_ch=64)))

2.$\mathbf{F}_1$=MaxPool2D($\mathbf{F}_0$,k=3,s=2,p=1)

3.for i=1 to N **do**// N **= 2**

z=ReLU(BN($Conv_{1\times1}$($\mathbf{F}_{1,i-1}$,out_ch=16)))

z=ReLU(BN($Conv_{3\times3}$(z,out_ch=16,stride=1)))

$\mathbf{F}_{1,i}$=BN($Conv_{1\times1}$(z,out_ch=64)))

4.$\mathbf{F}_2$ = MixTransformer(X, patch_size=16, in_chans=3, embed_dims=[64, 128, 256, 512],num _heads=[1, 2, 4, 8], mlp_ratios=[4, 4, 4, 4])//Based on segformer Transformer

5.$\mathbf{F}_3$ = concat($\mathbf{F}_{1,N}$,$\mathbf{F}_2$)

Algorithm 1. Parallel Block.Experiment results

This section will present the two datasets used in the experiments, the experimental environment, and the parameter settings. And displays our experimental results and results analysis.

Dataset

This study employs two datasets: one is the publicly available BioHit dataset, and the other is our self-constructed dataset.

(1) The public BioHit dataset comprises 300 tongue images, each with dimensions of 576×768 pixels. The ground truth for the data is based on manually annotated results.
(2) Our self-constructed dataset contains 642 photographs, collected by using the professional four-diagnostic device, with ground truth established through manual annotation.

We randomly divided the two tongue image datasets into training and testing sets in an 8:2 proportion. The dataset was annotated and partitioned into training, testing, and validation sets following the standard VOC dataset format. To ensure reproducible data splitting, The data was split using sklearn's train_test_split function with random seed set to 717,750, guaranteeing consistent data distribution across different experimental runs. Initially, all images were scaled down to a consistent resolution of 512 × 512 pixels. Subsequently, a horizontal flip augmentation was applied to the training images with a 50% probability to bolster the model's capacity for generalization and mitigate the risk of overfitting. Photometric distortions were also introduced to the images with a 50% probability of further augmenting the dataset. Additionally, the images in both the training and testing sets underwent standardization in the RGB channels.

Experimental environment and parameter settings

The deep learning framework employed in this study is PyTorch 1.11.0, paired with CUDA version 11.3 for GPU acceleration. In terms of hardware specifications, the setup includes an NVIDIA A4000 GPU featuring 16 GB of VRAM and an Intel Xeon(R) Gold 5320 CPU operating at a 2.20 GHz clock frequency. To ensure complete reproducibility of our experiments, we have fixed the random seed to 717,750 across all computational libraries including Python's random module, NumPy, and PyTorch. These measures guarantee that any variations in experimental results can be reliably attributed to model architecture or hyperparameter changes rather than computational randomness.

Model training method

In this study, the AdamW optimizer was chosen for the training of the enhanced network model, with an initial learning rate of 0.00006. The subtle updates to model weights throughout the training process facilitate a more nuanced convergence towards the minimum loss region. To counteract overfitting and bolster the model's generalization ability, the weight_decay parameter was set to 0.0001. The model was trained for a total of 100 epochs, with a mini-batch size set to 16. The Cross-Entropy loss function was leveraged to facilitate the optimization of the neural network's layer weights and biases.

Model evaluation indicators

To comprehensively evaluate the performance of PAPU_TonSeg, multiple evaluation indicators were adopted, including MPA (Mean Pixel Accuracy), MIoU (Mean Intersection over Union), the Dice coefficient, FLOPs (Floating Point Operations), and the parameter count of the model itself.

(1) MPA: The PA (Pixel Accuracy) can be obtained by calculating the proportion of the number of pixels that are correctly classified in each category to the total number of pixels in that category. Then, the MPA is obtained by averaging the PA values of all categories. The calculation formula of MA is presented in Formula 1. In this formula, n_{jj} is the total number of pixels marked as category j and correctly predicted as j, while t_j is the total number of pixels marked as j.

$$MPA = \frac{1}{k}\sum_{j=1}^{k}\frac{n_{jj}}{t_j} \tag{1}$$

(2) Dice: The Dice index is computed by evaluating the fraction of the intersection (i.e., the correctly identified pixel count) to the union (i.e., the sum of the predicted and actual pixel counts) of the predicted and true pixels. This index indicates the congruence between the model's predictive outcomes and the actual outcomes, hence acting as a significant criterion for gauging the model's segmentation accuracy. The Dice index varies from 0 to 1, with values tending toward 1 suggesting superior segmentation quality[37]. Equation 2 illustrates the method for calculating the Dice index, where TP indicates true positives, FP indicates false positives, and FN indicates false negatives[38].

$$Dice = \frac{2 * TP}{FP + 2 * TP + FN} \tag{2}$$

(3) MIoU: The MIoU (Mean Intersection over Union) is an indicator expanding upon IoU (Intersection over Union). IoU can quantitatively measure the overlap between predicted and actual regions. The MIoU method calculates IoU for each class separately and then computes the arithmetic average of these individual IoU values, which provides a comprehensive measure that accounts for all classes. This approach offers a holistic assessment of a model's performance in semantic segmentation. The formula for MIoU is detailed in Eq. 3 [39]. Here, *nii* signifies the aggregate pixel count designated as category *i* and accurately identified as such, whereas *nij* represents the pixel count for category *i* that has been misidentified as category *j*. The definitions for all other terms remain consistent with the aforementioned.

$$MIoU = \frac{TP}{TP + FP + FN} = \frac{1}{k}\sum_{i=0}^{k}\frac{n_{ii}}{\sum_{j=0}^{k} n_{ij} + \sum_{j=0}^{k} n_{ji} - n_{ii}} \tag{3}$$

Network	BioHit			Self-Built Dataset		
	MPA (%)	MIou (%)	Dice (%)	MPA (%)	MIou (%)	Dice (%)
FCN	91.75	94.92	92.32	90.30	95.02	93.86
UNet	84.33	92.76	86.55	94.94	97.33	96.80
PSPNet	93.96	97.35	94.95	95.10	96.78	97.00
DeepLabv3	**95.59**	**98.29**	**96.44**	95.83	97.93	97.41
Segformer	95.14	97.82	95.97	**96.68**	**98.31**	**97.92**
PAPU_TonSeg	**97.56**	**98.60**	**97.99**	**97.32**	**98.64**	**98.32**

Table 1. Comparative evaluation metrics of different models on the BioHit dataset and the self-built dataset. Significant values are in bold.

Network	FLOPs(G)	Param(M)
FCN	19.5	18.64
UNet	38.9	32.09
PSPNet	11.3	46.71
DeepLabv3	13.0	58.63
Segformer	1.3	3.71
PAPU_TonSeg	2.0	7.05

Table 2. FLOPs and param data for different models with the same input.

(4) FLOPs: As a pivotal indicator of the complexity inherent in algorithms or models, FLOPs (Floating Point Operations) are frequently applied to quantify the aggregate computational effort required by neural network models, thus offering an inference of the model's time complexity.

Experimental results

Comparison of each model

For a thorough assessment of model performance, the classic models mentioned above, the Segformer model, and the PAPU_TonSeg model were independently trained on both the publicly available BioHit dataset and our self-built dataset, followed by an evaluation of model performance on the respective test sets. The comparative performance of different models on both datasets is delineated in Table 1.

From Table 1, it can be observed that on the BioHit dataset, the performance of both DeepLabV3 and Segformer across all evaluation indicators is better than that of FCN, UNet, and PsPnet (Pyramid Scene Parsing Network) models. When comparing DeepLabV3 and Segformer, the metrics are quite close, with DeepLabV3 performing slightly better than Segformer. Among the five models, the improved PAPU_TonSeg model demonstrates the best performance. Compared to the original Segformer, the enhanced PAPU_TonSeg model shows a 2.42% increase in MPA, a 0.78% increase in Miou, and a 2.02% increase in Dice. When compared to DeepLabV3, the improved model exhibits a 1.97% increase in MPA, a 0.31% increase in MIOU, and a 1.55% increase in Dice.

On our self-constructed dataset, the original Segformer outperforms other classic models in all evaluation indicators, but the performance of the improved PAPU_TonSeg is further enhanced, becoming the best-performing model among all. The PAPU_TonSeg shows a 0.64% higher MIou than Segformer, and a 0.33% higher MIou, as well as a 0.4% higher Dice. Therefore, on both datasets, the PAPU_TonSeg model has the optimal effect, fully demonstrating the effectiveness of the improvement strategy and the robustness of the model.

To thoroughly evaluate the performance of different models, a comparison of FLOPs and Param was undertaken. Given the correlation between a model's FLOPs and the input data size, all models were subjected to data of uniform dimensions ($1 \times 3 \times 224 \times 224$). The Python third-party library thops was employed to ascertain the FLOPs for the different models, and the same tool was applied to determine the models' parameter quantities (Param). The experimental data are explicitly illustrated in Table 2.

Table 2 illustrates that the Segformer, in contrast to legacy semantic segmentation models that are predominantly CNN-based, possesses a considerably smaller parameter set and a significantly reduced computational complexity. The PAPU_TonSeg, after enhancements, sees an uptick of 0.7G in computational complexity and an addition of 3.34 M parameters over the original Segformer. Despite this, it remains on the lower end of the spectrum concerning parameter volume and computational complexity in comparison to other classic models.

To further compare the actual segmentation outcomes of various models, 4 random images were drawn from the custom dataset and the public BIOHit dataset for a comparative experiment of the actual segmentation performance of each model. The segmentation results of the models are shown in Figs. 4 and 5.

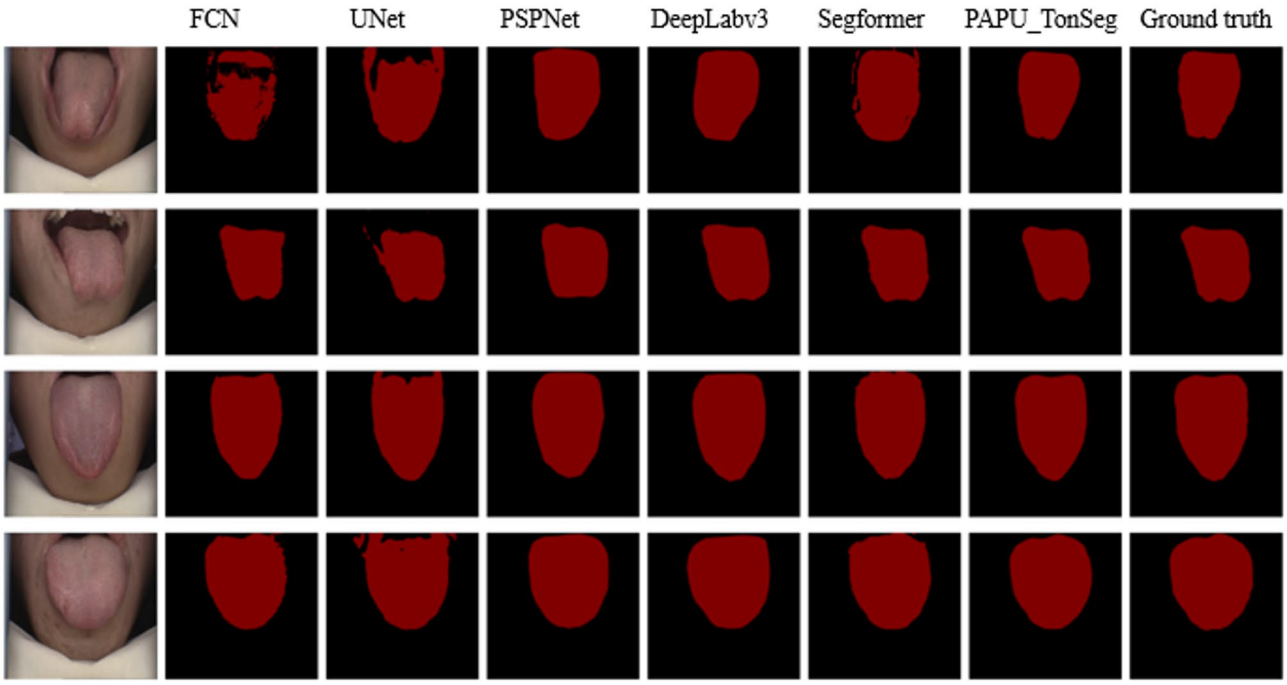

Fig. 4. Segmentation outcomes of different models on the public BioHit dataset.

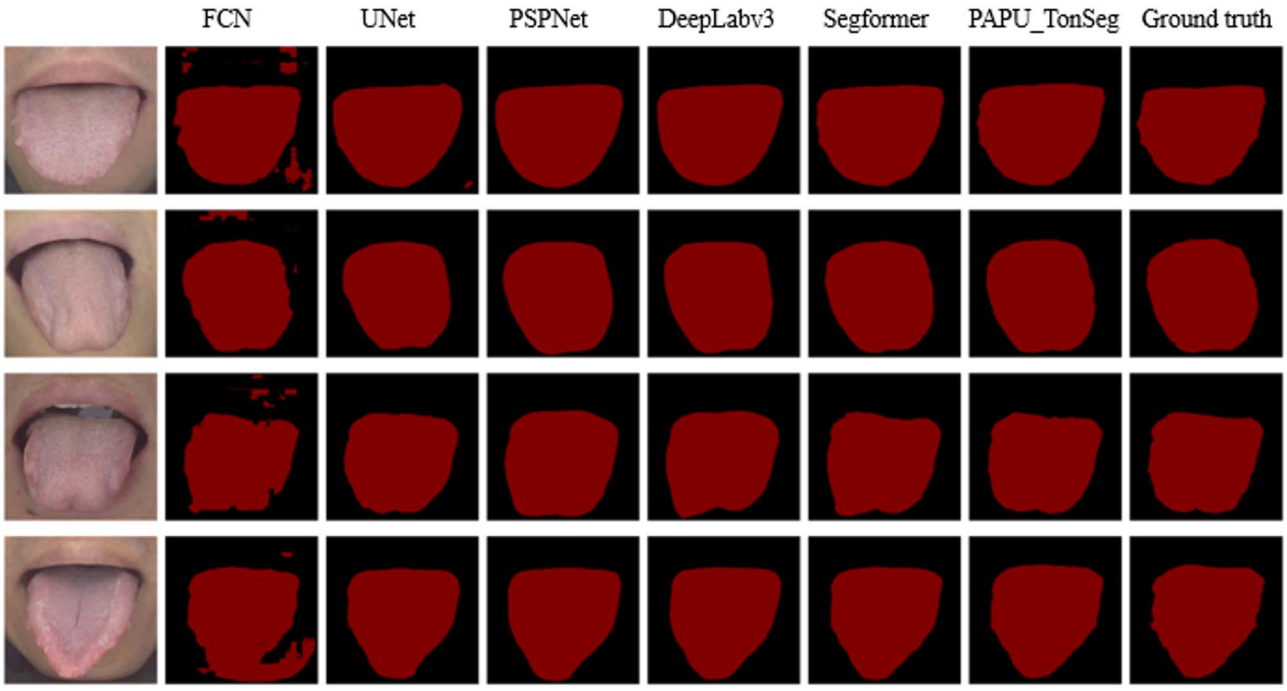

Fig. 5. Segmentation results of different models on the self-built dataset.

As shown in Fig. 4, the segmentation efficacy of four tongue images from the BioHit public dataset is presented. Each row corresponds to a sample, while each column displays the segmentation effect of a specific model on various samples. The sequence of models in the columns is Input, FCN, UNet, PSPNet, DeepLabV3, Segformer, PAPU_TonSeg, and Ground Truth. It can be observed that the segmentation effects of FCN and UNet models significantly differ from Ground Truth, with torn segmentation and blurred edges; PSPNet model's segmentation, while similar to Ground Truth, lacks edge clarity; UNet, DeepLabV3, and Segformer also exhibit a lack of refinement in segmenting tongue body details and tooth marks. Compared to other models, the improved PAPU_TonSeg model's segmentation outcome is the closest to Ground Truth, which effectively distinguishes the

tongue body without debris, and accurately segments the tooth marks on the tongue body edge, resulting in better segmentation detail of the tongue body edge.

Figure 5 illustrates the comparative segmentation outcomes of various models on our dataset. Upon examination of the segmentation results, it is evident that the enhanced model effectively delineates edge details, such as tooth marks, etc. Despite PSPNet, DeepLabV3, and Segformer capturing the general contour of the tongue body, they fail to provide crisp segmentation of the edges, demonstrating a marked improvement in edge detail segmentation accuracy with our refined PAPU_TonSeg.

Ablation experiment

Building upon the foundational framework of the Segformer model, we implemented critical enhancements by replacing the Transformer Block in the Segformer's backbone network with a Parallel block, infusing the Mix-FFN with the ECA attention mechanism, and enhancing the Segformer's up-sampling procedure with the Multi-Level Attention (MLA) module. To substantiate the effectiveness of these refinements, comprehensive ablation experiments were executed on the BioHit public dataset, with the employment of Mean Pixel Accuracy (MPA) as the evaluative criterion for model performance. The tabulated experimental findings are illustrated in Table 3.

Feature visualization

The Grad-Cam technique (Gradient-weighted Class Activation Mapping) serves to demystify the decision processes of Convolutional Neural Networks. By analyzing the gradients, it identifies the most impactful areas of the input image on the network's predictions[40–43].

Figure 6 illustrates the heatmaps obtained from applying Grad-Cam to three randomly selected images each from the public dataset BioHit and our in-house dataset.

As depicted in Fig. 6, the original images are presented on the extreme left, followed by heatmaps generated from the FCN, UNet, PSPNet, DeepLabV3, Segformer, and PAPU_TonSeg model moving from left to right. The FCN and UNet are observed to target only particular tongue local features, bypassing a substantial portion of the tongue's informative features. The UNet model even notably misidentifies regions such as the lips as focus areas. The PSPNet model intermittently does not fully capture the tongue's edge features, and the DeepLabV3 and Segformer also demonstrate an incomplete recognition of the tongue with an undue focus on the tongue's central regions. However, the refined PAPU_TonSeg model's attention scope encompasses the entire tongue, and the heatmap color distribution reveals a more consistent focus area, with edges and the center showing nearly identical color saturation. This indicates that the modified model effectively acquires both global and local features of the tongue, facilitating accurate segmentation based on tongue attributes.

Comparative analysis with general models

To demonstrate the superior performance of our improved model in tongue image segmentation, we conducted comparative experiments with state-of-the-art models in recent years. All models were trained and evaluated within the mmsegmentation framework. Specifically, we adapted our new model architecture to conform to mmsegmentation's standard structure and jointly trained it using both our proprietary dataset and the BioHit dataset under identical training parameters.For optimization, all models employed the SGD optimizer with a learning rate of 0.01 and a momentum of 0.9. As shown in Table 4, the improved model exhibits significantly better performance compared to general-purpose models, indicating that our proposed enhancements enable superior applicability in tongue image segmentation tasks.

Convergence analysis

To demonstrate the variation of the loss function during the training process, thereby verifying the stability of the model and the absence of overfitting, we plotted the loss curve of the model on the validation set. As shown in the Fig. 7, the loss exhibits significant fluctuations in the early stages of training. However, in the later stages, the loss function on the validation set gradually stabilizes. Moreover, during the latter half of the training process, the validation loss does not show any substantial upward trend, indicating that the model did not exhibit overfitting during training.

Discussion

Tongue diagnosis can enable doctors to reveal the health condition of the body's inner organs by observing changes in the tongue motility and coating. However, the diagnostic method of the tongue is easily influenced by the physician's professional knowledge, personal experience, and external environmental factors, which leads

Network	Parallel	Eca attention	Progressive upsample	MPA(%)
Segformer				95.14
Segformer + Parallele	√			97.25
Segformer + Eca Attention		√		97.06
Segformer + Progressive Upsample			√	96.80
Segformer + Parallel + Eca Attention	√	√		97.33
PAPU_TonSeg	√	√	√	98.60

Table 3. Results of ablation study on the public dataset BioHit.

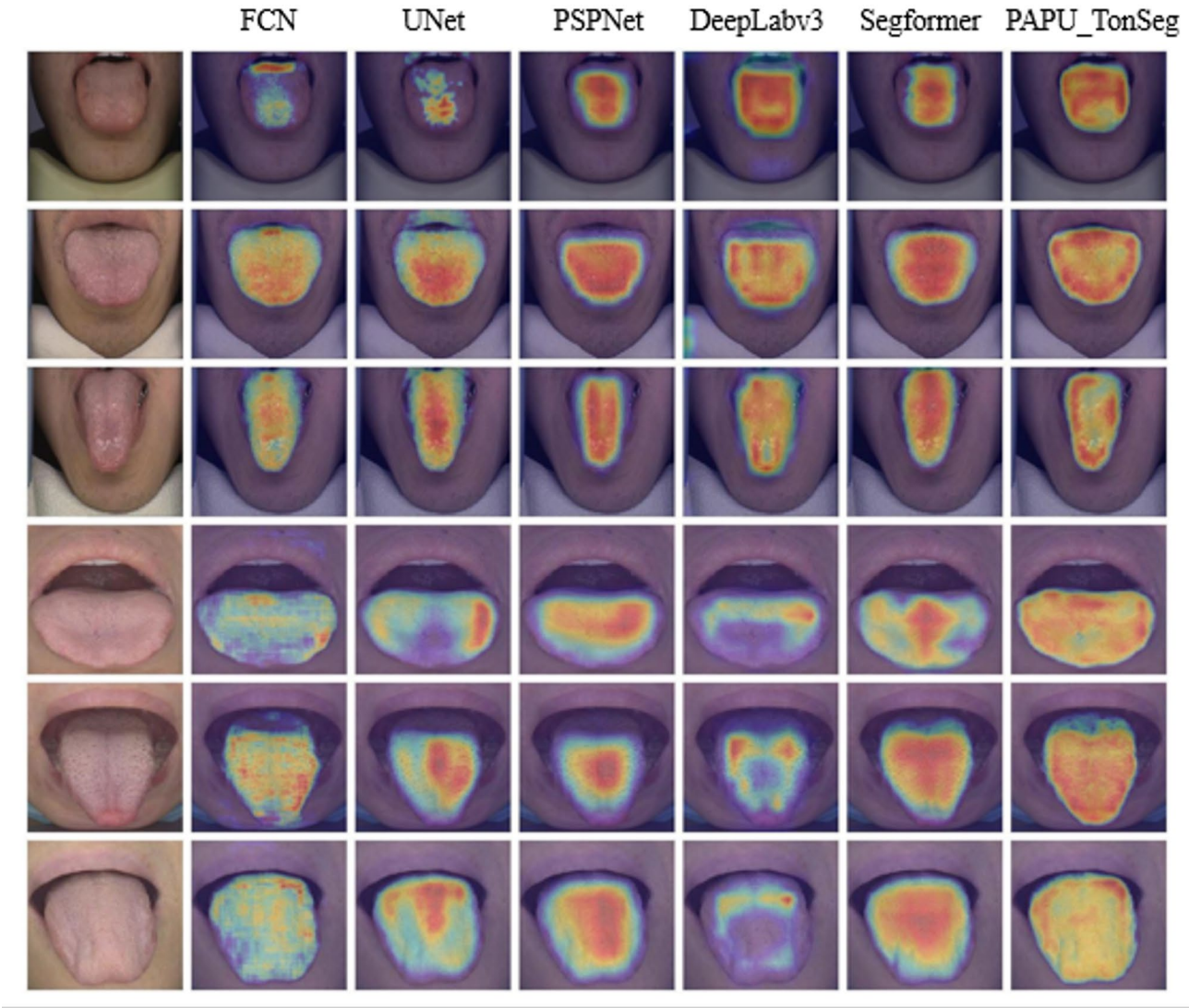

Fig. 6. Grad-Cam Heatmaps of different models on the BioHit dataset and our self-built dataset.

	IoU	Acc	Dice	Precision	Recall
Vit	93.23	96.73	96.50	96.27	96.73
HRNet	94.12	95.99	96.97	97.97	95.99
Swin Transformer	95.80	97.56	97.86	98.15	97.56
Beit	91.77	92.79	95.71	93.72	97.79
PAPU_TonSeg	96.82	98.21	98.38	98.55	98.21

Table 4. Comparison with baseline models.

to subjectivity to some degree. To boost the objectivity and precision of TCM tongue diagnosis and elevate the automation and intelligent capabilities of TCM, the implementation of AI-driven TCM tongue diagnosis techniques is crucial. For the tongue image classification models to ascertain diagnoses accurately, precise and exhaustive capture of the patient's tongue imagery is fundamental, rendering a superior tongue image semantic segmentation model critical.

Our study proposes the PAPU_TonSeg, an improvement based on the lightweight Segformer model. To begin with, the integration of a parallel network architecture that encompasses self-attention and residual components within the original Segformer allows for the simultaneous extraction of global and local information. Subsequently, the incorporation of the ECA attention mechanism into the Mix-FFN framework significantly enhances the network's ability to extract features, improving the accuracy and efficiency of semantic segmentation. Finally, the utilization of a progressive upsampling approach in conjunction with a convolutional

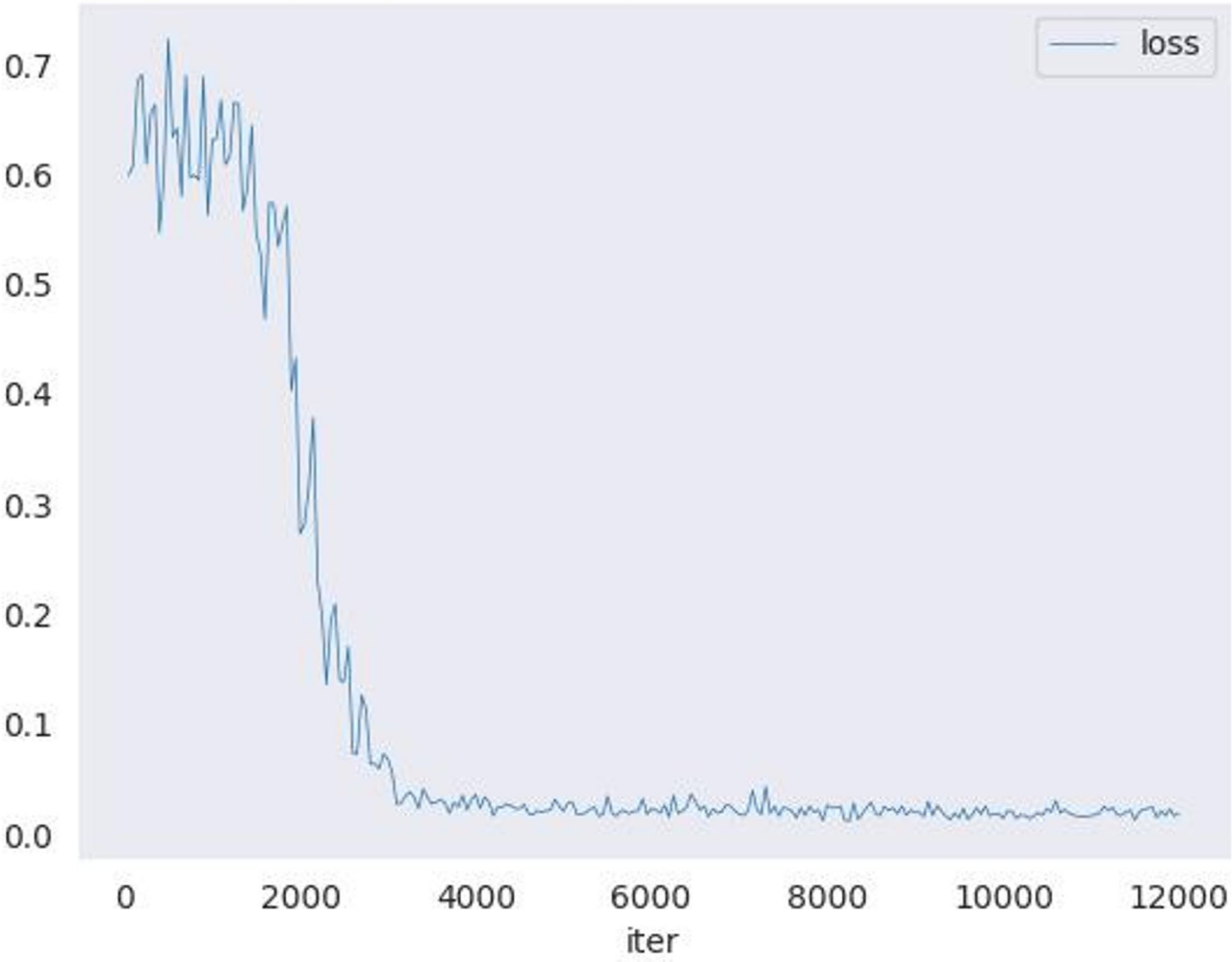

Fig. 7. Loss curve of the PAPU_TonSeg model on the validation set.

network model minimizes information loss during upsampling, ensuring that more edge detail is retained throughout the upsampling process.

In this research, despite a few attempted enhancements to the Segformer model aimed at boosting the efficacy of tongue image segmentation, we faced several challenges and identified areas for improvement. (1) During the conduct of ablation tests, each enhancement did result in a notable performance improvement compared with the original model, however, the segmentation performance has not achieved the most desirable outcomes, suggesting that there remains scope for enhancement in the indicators used for practical evaluation. (2) The existing evaluation frameworks have shown inadequacies, particularly in evaluating the precision of edge segmentation. These shortcomings are significant obstacles to be tackled in future research endeavors.

Conclusion

This study made a series of improvements based on the original Segformer to address the existing issues in tongue body segmentation. The experimental outcomes indicate that these enhancements have elevated the model's segmentation capabilities across two tongue image datasets, which outperforms legacy models. The PAPU_TonSeg, which we have refined, exhibits reduced computational complexity compared to conventional convolutional neural network-based semantic segmentation models, and it more precisely delineates the tongue body, notably resolving the intricate segmentation of tongue body edge details. This study endeavor contributes positively to the scientize of tongue diagnosis, heightening its objectivity and exactitude, and possesses appreciable value for wider dissemination and practical use.

Data availability

The public BioHit dataset: https://github.com/BioHit/TongeImageDataset.

Received: 11 February 2025; Accepted: 2 June 2025

Published online: 29 July 2025

References

1. Hsu, P. C. et al. The association between arterial stiffness and tongue manifestations of blood stasis in patients with type 2 diabetes. *BMC Complement Altern. Med.* **16**, 324. https://doi.org/10.1186/s12906-016-1308-5 (2016).
2. Han, S. et al. Potential screening and early diagnosis method for cancer: Tongue diagnosis. *Int. J. Oncol.* **48**(6), 2257–2264. https://doi.org/10.3892/ijo.2016.3466 (2016).
3. Ni, J., Yan, Z. & Jiang, J. Tongue caps: An improved capsule network model for multi-classification of tongue color. *Diagnostics* **12**(3), 653. https://doi.org/10.3390/diagnostics12030653 (2022).
4. Pang, B., Zhang, D., Li, N. & Wang, K. Computerized tongue diagnosis based on Bayesian networks. *IEEE Trans. Biomed. Eng.* **51**(10), 1803–1810. https://doi.org/10.1109/TBME.2004.831534 (2004).
5. Pang, B., Zhang, D. & Wang, K. The bi-elliptical deformable contour and its application to automated tongue segmentation in Chinese medicine. *IEEE Trans. Med. Imaging* **24**(8), 946–956. https://doi.org/10.1109/TMI.2005.850552 (2005).
6. Shi, M., Li, G. & Li, F. C^2G^2FSnake: Automatic tongue image segmentation utilizing prior knowledge. *Sci. China Inf. Sci.* **56**, 1–14. https://doi.org/10.1007/s11432-011-4428-z (2013).
7. Shi, M. J., Li, G. Z., Li, F. F. & Xu, C. Computerized tongue image segmentation via the double geo-vector flow. *Chin Med.* **9**(1), 7. https://doi.org/10.1186/1749-8546-9-7 (2014).
8. Yan, B., Zhang, S., Yang, Z., Su, H. & Zheng, H. Tongue segmentation and color classification using deep convolutional neural networks. *Mathematics* **10**(22), 4286. https://doi.org/10.3390/math10224286 (2022).
9. "Flaws Can Be Applause: Unleashing Potential of Segmenting Ambiguous Objects in SAM | OpenReview"; accessed 11 May 2025. https://openreview.net/forum?id=vJSNsSFO95 (n.d.)
10. "P^2SAM: Probabilistically Prompted SAMs Are Efficient Segmentator for Ambiguous Medical Images | OpenReview."; accessed 11 May 2025. https://openreview.net/forum?id=IjULktl95t (n.d.).
11. Li, C. et al. GTP-4o: Modality-prompted heterogeneous graph learning for Omni-modal biomedical representation (2024). https://doi.org/10.48550/arXiv.2407.05540.
12. Sun, L. et al. Few-shot medical image segmentation using a global correlation network with discriminative embedding. *Comput. Biol. Med.* **140**(January), 105067. https://doi.org/10.1016/j.compbiomed.2021.105067 (2022).
13. Lopez Pinaya, W. H. Unsupervised Brain Anomaly Detection and Segmentation with Transformers. https://doi.org/10.48550/arXiv.2102.11650 (2021).
14. Zhang, Y. Generator versus Segmentor: Pseudo-healthy synthesis. https://doi.org/10.48550/arXiv.2009.05722 (2021).
15. Long, J., Shelhamer, E. & Darrell, T. Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **4**(39), 640–651. https://doi.org/10.1109/TPAMI.2016.2572683 (2017).
16. Xu, H. et al. A two-stage segmentation of sublingual veins based on compact fully convolutional networks for traditional Chinese medicine images. *Health Inf. Sci. Syst.* **11**, 19. https://doi.org/10.1007/s13755-023-00214-1 (2023).
17. Zhou, C., Fan, H. & Li, Z. Tonguenet: Accurate localization and segmentation for tongue images using deep neural networks. *IEEE Access* **7**, 148779–148789. https://doi.org/10.1109/ACCESS.2019.2946681 (2019).
18. Ronneberger, O., Fischer, P. & Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Springer, 2015). https://doi.org/10.1007/978-3-319-24574-4_28.
19. Zunair, H. & Hamza, A. B. Sharp U-Net: Depthwise convolutional network for biomedical image segmentation. *Comput. Biol. Med.* **136**, 104699. https://doi.org/10.1016/j.compbiomed.2021.104699 (2021).
20. Xu, Q. et al. Multi-task joint learning model for segmenting and classifying tongue images using a deep neural network. *IEEE J. Biomed. Health Inform.* **24**(9), 2481–2489. https://doi.org/10.1109/JBHI.2020.2986376 (2020).
21. Li, M. Y. et al. Application of U-Net with global convolution network module in computer-aided tongue diagnosis. *J. Healthc. Eng.* **2021**, 5853128. https://doi.org/10.1155/2021/5853128 (2021).
22. Chen, S., Gamechi, Z. S., Dubost, F., van Tulder, G. & de Bruijne, M. An end-to-end approach to segmentation in medical images with CNN and posterior-CRF. *Med. Image Anal.* **76**, 102311. https://doi.org/10.1016/j.media.2021.102311 (2022).
23. Zhang, X. et al. An improved tongue image segmentation algorithm based on the DeepLabv3+ framework. *IET Image Proc.* **16**(5), 1473–1485. https://doi.org/10.1049/ipr2.12425 (2022).
24. Liu, H. et al. MEA-Net: Multilayer edge attention network for medical image segmentation. *Sci. Rep.* **12**(1), 7868. https://doi.org/10.1038/s41598-022-11852-y (2022).
25. Yang, J., Tu, J., Zhang, X., Yu, S. & Zheng, X. TSE DeepLab: An efficient visual transformer for medical image segmentation. *Biomed. Signal Process. Control* **80**(Part 2), 104376. https://doi.org/10.1016/j.bspc.2022.104376 (2023).
26. Chen, L. C. et al. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: *Proceedings of the European Conference on Computer Vision (ECCV)* 801–818 (2018).
27. Zhang, X. et al. An improved tongue image segmentation algorithm based on the Deeplabv3+ framework. *IET Image Process*. **16**. https://doi.org/10.1049/ipr2.12425 (2022).
28. Vaswani, A. et al. Attention is all you need. https://doi.org/10.48550/arXiv.1706.03762 (2017).
29. Dosovitskiy, A. et al. An image is worth 16x16 words: Transformers for image recognition at scale. http://arxiv.org/abs/2010.11929 (2021).
30. Xie, E. et al. Segformer: Simple and efficient design for semantic segmentation with transformers. *Adv. Neural Inf. Process.* **34**, 12077–12090. https://doi.org/10.48550/arXiv.2105.15203 (2021).
31. Szegedy, C. et al. Inception-v4, Inception-ResNet and the impact of residual connections on learning. http://arxiv.org/abs/1602.07261 (2016).
32. Iosifidis, A. & Tefas, A. (eds) *Deep Learning for Robot Perception and Cognition* (Academic Press, 2022).
33. Li, C. et al. U-Kan makes strong backbone for medical image segmentation and generation. https://arxiv.org/abs/2406.02918 (2024).
34. Li, C. et al. Hierarchical deep network with uncertainty-aware semi-supervised learning for vessel segmentation. https://arxiv.org/abs/2105.14732 (2021).
35. Oktay, O. et al. Attention U-Net: Learning where to look for the pancreas. https://arxiv.org/abs/1804.03999 (2018).
36. Hu, J. et al. Sa-Net: A scale-attention network for medical image segmentation. PloS One. Accessed 31 March 2025. https://pubmed.ncbi.nlm.nih.gov/33852577/.
37. Hu, J. et al. Squeeze-and-excitation networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 99. https://doi.org/10.1109/TPAMI.2019.2913372 (2017).
38. Wang, Q. et al. ECA-Net: Efficient channel attention for deep convolutional neural networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2020). https://doi.org/10.1109/CVPR42600.2020.01155.
39. Yan, J. et al. Tongue crack recognition using segmentation-based deep learning. *Sci. Rep.* **13**, 511. https://doi.org/10.1038/s41598-022-27210-x (2023).
40. Shamir, R. R. et al. Continuous dice coefficient: A method for evaluating Probab29ilistic segmentations (Cold Spring Harbor Laboratory, 2018). https://doi.org/10.1101/306977.

41. Jiang, H. et al. A review of deep learning-based multiple-lesion recognition from medical images: Classification, detection and segmentation. *Comput. Biol. Med.* **157**, 106726. https://doi.org/10.1016/j.compbiomed.2023.106726 (2023).
42. Li, J. et al. Automatic classification framework of tongue feature based on convolutional neural networks. *Micromachines* **13**(4), 501. https://doi.org/10.3390/mi13040501 (2022).
43. Selvaraju, R. R. et al. Grad-CAM: Visual explanations from deep networks via gradient-based localization. *Int. J. Comput. Vision* **128**(2), 336–359. https://doi.org/10.1007/s11263-019-01228-7 (2020).

Acknowledgements

This work was supported by the horizontal project of Beijing University of Chinese Medicine (BUCM-2024-JS-FW-137).

Author contributions

All authors contributed to the study conception and design. X.W.: Writing, Software, Methodology. Y.C.: Writing, review & editing. Y.C.: Writing—review & editing. H.L.: Resources, Material preparation and data collection. A.H.: Experimental Methods, Model design. Y.T: Funding acquisition, Experimental Methods, Model design.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to H.L., A.H. or Y.T.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.