Original Research Article

DIGITAL HEALTH

DIGITAL HEALTH
Volume 10: 1–14

DOI: 10.1177/20552076241289007
journals.sagepub.com/home/dhj

Sage

TUMamba: A novel tongue segment methods based on Mamba and U-Net

Fan Jiang[1], Yanmei Zhong[2] and Simin Yang[3]

Abstract

Background and Objective: Current tongue segmentation methods often struggle with extracting global features and performing selective filtering, particularly in complex environments where background objects resemble the tongue. These challenges significantly reduce segmentation efficiency. To address these issues, this article proposes a novel model for tongue segmentation in complex environments, combining Mamba and U-Net. By leveraging Mamba's global feature selection capabilities, this model assists U-Net in accurately excluding tongue-like objects from the background, thereby enhancing segmentation accuracy and efficiency.

Methods: To improved the segmentation accuracy of the U-Net backbone model, we incorporated the Mamba attention module along with a multi-stage feature fusion module. The Mamba attention module serially connects spatial and channel attention mechanisms at the U-Net 's skip connections, selectively filtering the feature maps passed into the deep network. Additionally, the multi-stage feature fusion module integrates feature maps from different stages, further improving segmentation performance.

Results: Compared with state-of-the-art semantic segmentation and tongue segmentation models, our model improved the mean intersection over union by 1.17%. Ablation experiments further demonstrated that each module proposed in this study contributes to enhancing the model's segmentation efficiency.

Conclusion: This study constructs a **T**ongue segmentation model based on **U**-Net and **Mamba** (TUMamba). The model effectively extracted global spatial and channel features using the Mamba attention module, captured local detail features through U-Net, and enhanced image features via multi-stage feature fusion. The results demonstrate that the model performs exceptionally well in tongue segmentation tasks, proving its value in handling complex environments.

Keywords

Intelligent tongue segmentation, Mamba, U-Net, multi-stage feature fusion

Submission date: 9 April 2024; Acceptance date: 10 September 2024

Introduction

As an integral part of traditional medicine, traditional Chinese medicine (TCM) is gaining increasing global recognition.[1,2] In China, TCM has demonstrated remarkable efficacy in treating chronic ailments such as diabetes,[3] gastritis,[4] insomnia,[5] etc., which can not be separated from its unique diagnostic approach and reliable clinical practice. One of the key diagnostic methods in TCM is observation, which alongside smelling and listening, inquiring, pulse feeling, and palpation, forms the basis of diagnosis.

[1]Rehabilitation Medicine Department, The first Affiliated Hospital of Chongqing Medical University, Chongqing, China
[2]School of Foreign Languages and Cultures, Chongqing University, Chongqing, China
[3]Rehabilitation Medicine Department, Sichuan Jiaotong Hospital, Chengdu, China

Corresponding author:
Simin Yang, Sichuan Jiaotong Hospital, 139 North Hengshan Street, Chengdu, Sichuan 611730, China.
Email: siminyang0326@163.com

Among these, observation has garnered significant attention in the field of medical image processing, especially in intelligent tongue diagnosis, due to its intuitive nature[6–8]; therefore, tongue diagnosis is extensively researched because it reflects a wealth of information about the human body and is relatively easy to collect data.[9]

In the context of intelligent tongue diagnosis, accurate tongue segmentation plays a critical role in achieving precise diagnostic outcomes. Despite the considerable research conducted on tongue segmentation[10] challenges persist, particularly as intelligent tongue diagnosis becomes increasingly utilized on mobile devices. Existing methods often result in mis-segmentation or over-segmentation due to factors such as skin variations and environmental conditions. This underscores the necessity for a tongue segmentation model that can adapt to complex environments with greater accuracy.[11] For example, U-Net has been favored in the field of tongue segmentation due to its compact size and strong ability to extract local features. However, its unique U-shaped structure and convolutional neural network (CNN)-based architecture present certain challenges. While the U-shaped structure facilitates the fusion of shallow and deep image features, the direct skip connections can inadvertently introduce shallow noise into deeper layers. To mitigate this issue, attention blocks have been incorporated into skip connections as a promising method for filtering out shallow noise.[12,13] However, these attention blocks are prone to over-fitting in complex environments where tongue images are limited, even though they significantly improve the model's inference time.[14] Moreover, while CNN-based models excel at extracting local features, they often struggle to capture global features, which is crucial for accurately segmenting tongue-like objects in challenging environments. This difficulty is exacerbated by noise transfer through skip connections and the model's inherent limitations in global feature extraction. Given these challenges, it is urgent to develop a robust noise filtering method for skip connections that can also effectively extract global features and select relevant features for transmission. Manifestly, the Mamba model offers a potential solution by converting images into text sequences, thereby enhancing global feature extraction. Additionally, it incorporates a feature selection mechanism that selectively retains or discards information based on input content,[15] aligning with the objectives of this study. Therefore, we proposed an innovative feature filtering and extraction approach, inspired by the convolutional block attention module (CBAM)[16] and based on Mamba, for the skip connection in U-Net. This method effectively extracts global features while selectively preserving spatial and channel features. Furthermore, to enhance feature representation, we employed a fusion method that integrates the output features from each stage, thereby improving the model's ability to capture local detail features.

The main contributions of this article are as follows:

1. This study is the first, to our knowledge, to leverage the capabilities of Mamba for global feature extraction and feature selection. This study could address the problem of U-Net's limited ability to extract global features and skip-connection's inability to select passing features.
2. To bolster U-Net's local feature extraction capability, we proposed a multi-stage feature fusion approach. This enhanced the local features extracted by the model.
3. Our model achieved superior results in the tongue segmentation task, proving its efficacy for the complex environment tongue segmentation task. This could provide a strong foundation for subsequent intelligent tongue diagnosis in similarly challenging environments.

The rest of this article is organized as follows. The "Related work" section provides a review of the related work. The "Materials and methods" section presents the new model proposed in this article. The "Results" section presents the experiment results. The "Discussion" section analyzes the experimental results. Finally, the conclusion of this article is presented in the "Conclusion" section.

Related work

Tongue segmentation

Existing tongue segmentation methods can be categorized into three main approaches: threshold segmentation methods, active contour methods, and deep learning methods. As one of the earliest techniques used for tongue segmentation, threshold segmentation relies on the color space features of the tongue and background in an image, setting threshold values to distinguish between them. For instance, Du et al.[17] converted red, green, blue tongue images to hue, saturation, and intensity (HSI) color space, using thresholds based on hue and intensity components to segment the tongue, followed by converting the images into binary form. Similarly, Chen et al.[18] employed a chroma threshold in hue, saturation, and value space, applying color enhancement in the region of interest and then using luminance thresholds in LAB color space to obtain the complete contour. Huang et al.[19] also converted tongue images to HSI color space, analyzing hue, and intensity characteristics of neighboring sub-blocks based on the user-selected seed points to merge sub-blocks that met the criteria. However, threshold segmentation becomes less effective in complex environments where parts of the image, such as skin, resemble the tongue, leading to inaccurate results. To overcome this, template matching segmentation was introduced, constructing templates to compare similar regions in images. For example, Saparudin and Muhammad[20] and Gu et al.[21] employed

threshold and Canny algorithms to detect tongue edges, followed by boundary refinement using Chan-Vese modeling. Despite these advancements, these methods struggle in complex environments where tongue-like noise significantly reduce segmentation performance. To address these challenges, deep learning methods have emerged, offering superior performance by effectively capturing semantic information in images and reducing the impact of noise. Among these, U-Net has become the backbone network for many tongue segmentation models.[22–24] However, with the growing demand for mobile tongue diagnosis, recent studies have focused on developing smaller and faster networks that can still accurately segment the tongue.[25–27] Yet, these models often compromise performance due to environmental noise, leading to incomplete segmentation. Given the critical importance of accurately segmenting the tongue, especially for analyzing tongue coating, the priority has shifted from reducing model complexity to improving segmentation accuracy in complex environments.

Attention mechanism

In segmentation models, the attention mechanism directs the model's focus to the target region. To extract spatial and channel features from images, some studies have incorporated channel attention and spatial attention into models, particularly for medical images such as computed tomography,[28] magnetic resonance imaging,[29] tongue diagnostic,[30] and cell images.[31] However, using spatial or channel attention alone in datasets with limited samples often fails to distinguish between features that are similar to the target, leading to over-fitting. To address this issue, integrated methods like CBAM and coordinated attention, which combine spatial and channel attention, have been developed.[32–34] Two common approaches have emerged for combining these attention features: parallel and serial connections of spatial and channel attention modules. The parallel approach, exemplified by DA-Net,[35] quickly extracts various attention features, enhancing the model's efficiency. In contrast, the serial connection, represented by CBAM.[16] allows for multiple extractions of attention features from the same data, yielding deeper semantic features.[36–38] However, feature loss during extraction can reduce the efficiency of backward feature extraction, impacting model performance. With the advancements in deep learning and the availability of enriched datasets, deeper networks like the transformer, based on the multi-head self-attention mechanism, have gained popularity in medical image processing.[39–41] Unlike CNNs, transformers can capture global features by dividing an image into smaller patches and processing them as a sequence, similar to a sentence. However, transformers lack inductive bias and require large datasets to achieve optimal results. To overcome these limitations, recent studies have explored combining CNNs with transformers to acquire both local and global features.[42–44] Although this hybrid approach addresses some of the transformer's shortcomings, it has not significantly improved performance on small datasets. Moreover, transformers are limited in mining long-range dependencies within images due to word size constraints, and their numerous parameters result in time-consuming inference.

State-space model

To address the challenges posed by the lengthy inference times of transformers and the difficulties CNNs face in capturing global features, the sequence-structured state-space model was proposed by Gu et al.[45] To further enhance model inference speed, Jimmy et al.[46] introduced S5, which reduces model complexity to a linear level. Subsequently, Harsh et al.[47] developed H3, a model that uses a structure similar to self-attention, making the feature extraction capabilities of the sequence-structured state-space model comparable to that of the transformer. The Mamba model, incorporating a gating structure and H3, along with an input adaptive mechanism, further enhances the state-space model, improving inference speed, throughput, and overall metrics.[15] As Mamba's effectiveness in processing continuous data and text data becomes evident, similar to the Transformer, research has increasingly applied it to image processing. Gu et al.[15] proposed a Mamba-based bi-directional positional coding module for image feature extraction, outperforming established vision transformer methods and significantly improving computational and memory efficiency. Zhu et al.[48] introduced a method to extract positional information of a two-dimensional (2D) image in four directions and connected the Mamba modules in series. Recently, an increasing number of researchers have applied visual Mamba, based on the multi-directional extraction of image position information, to medical image processing, where each Mamba module is combined in a U-shaped structure.[49] Although these methods integrate shallow features of structured spatial states with deep features through a U-shaped structure similar to U-Net, they do not effectively prevent noise in the shallow features from being transmitted to the deep features. One of the advantages of spatial state models is feature selection, which is also required in U-Net skip connections. However, there is still no effective method to select the transferring features for U-Net using this property. Similar to transformers, existing Mamba-based computer vision models lack effective local feature extraction methods and typically consist of stacked Mamba blocks, which also require pre-training in data-scarce medical image processing tasks. Many models have adopted the 2D-selective-scan (SS2D) method to assist the spatial state model in acquiring image spatial state features, but have not focused on channel features. Research in attention mechanisms suggests that the fusion

of channel and spatial features can further enhance a model's image-processing capabilities. Therefore, integrating channel and spatial feature selection methods to aid U-Net skip connections in filtering noise through the spatial state model's feature selection property holds significant potential.

Materials and methods

The structure of the proposed TUMamba is shown in Figure 1. The model's backbone is U-Net, and to mitigate the noise potentially introduced by skip connections, a Mamba-based approach was employed. This approach constructs an attention-like module that filters noise in both spatial and channel dimensions. Additionally, to enhance the model's ability in extracting detailed features, we fused the outputs of each stage. The loss was computed on the output of each stage to improve the preservation of deep feature details. This section details the Mamba attention module and the multi-stage feature fusion module that we developed.

Mamba attention module

In U-shaped networks, shallow spatial and channel features are often transmitted through skip connections to deeper layers. While this direct transmission helps retain useful spatial and channel information for segmentation, it also carries noise from the shallow features. For example, in the initial stage of a U-shaped network, direct transmission features can introduce up to 50% noise to the features

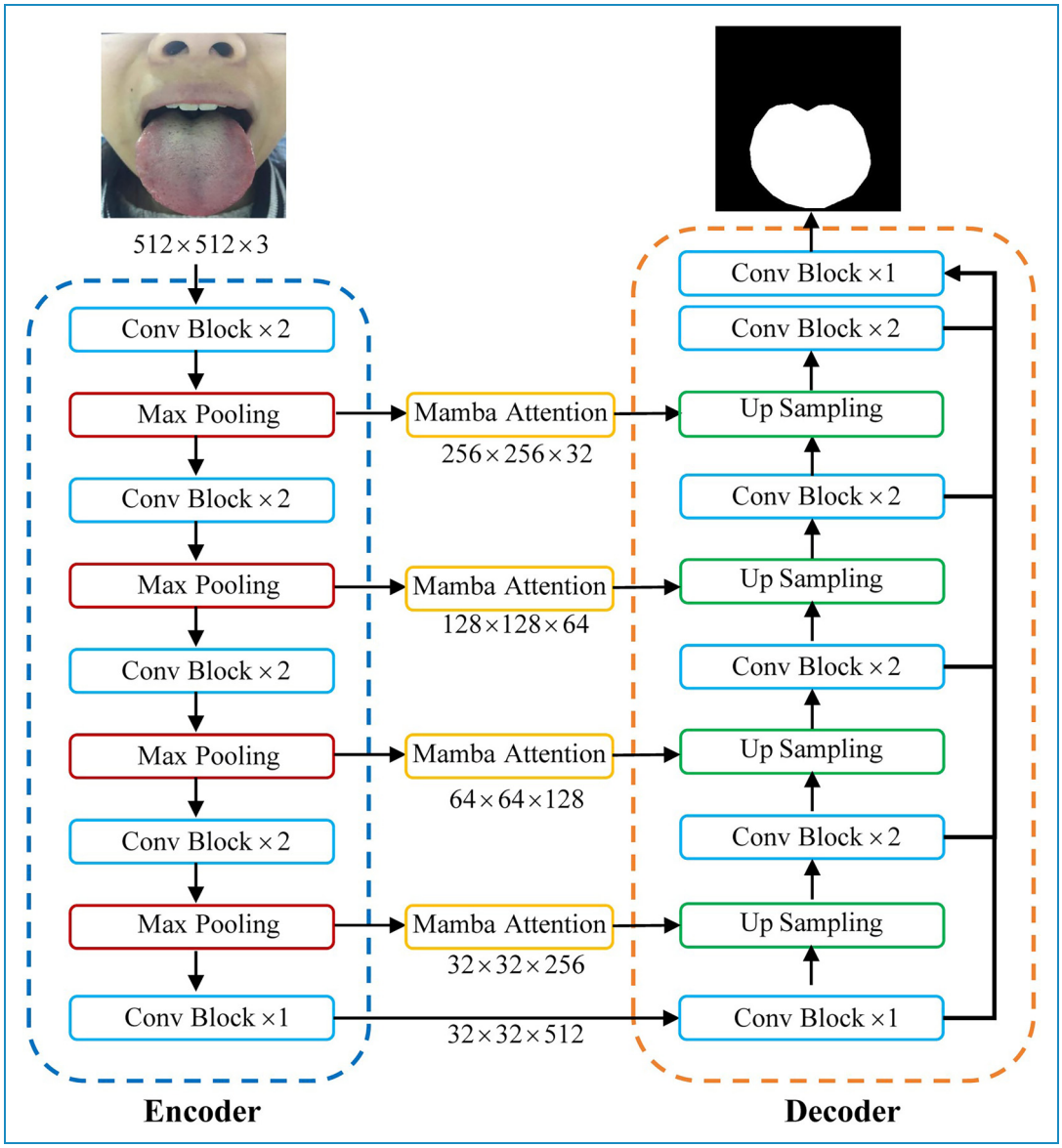

Figure 1. Overview of the model.

before convolution. Reducing this noise is crucial, particularly when segmenting the tongue in complex environments. Thus, replacing direct skip connection with a noise-reducing module that enhances shallow spatial and channel characteristics is necessary. Previous studies have indicated that attention mechanisms for tongue segmentation can lead to over-segmentation in complex environments due to small sample sizes and dynamic noise.[11,12] To address these challenges, we propose the Mamba attention module, which integrates the selective state space model into attention extraction. Similar to CBAM, this module serially extracts spatial and channel features from tongue images. The Mamba attention module consists of patch embedding, spatial Mamba attention, and channel Mamba attention, as shown in Figure 2.

To process images using Mamba, they are first subjected to patch embedding, similar to the vision transformer.[48] The image is divided into patches of size $N * N$. Next, get M splitted patches of size $N * N$. Assuming that the size of the image is $H * W * C$, then M here is $H * W * C/N * N$. A sequence $\mathbf{X}$ of shape $B * (N * N) * M$ is obtained, where B is the batch size. This sequence is then embedded to extract the image features after patch embedding. The specific method is shown in equation (1).

$$\mathbf{E} = \mathbf{Embedding}(\mathbf{X}) \tag{1}$$

Compared to CNN, Mamba excels in mining the long-range sequences information through the state-space model, offering faster computation and fewer parameters than transformer. However, as Mamba operates on sequence features, converting the image into a sequence first when dealing with 2D images. This operation causes the data to lose spatial information. Therefore, when processing the sequence consisting of patches, the spatial information of the 2D image is preserved by dividing the transformed features into four sets of sequences with different orders.[49] These four sequences are shown in Figure 3, which are S-shaped front-to-back, S-shaped back-to-front, N-shaped front-to-back, and N-shaped back-to-front. This ensures that each element in the feature map incorporates elements

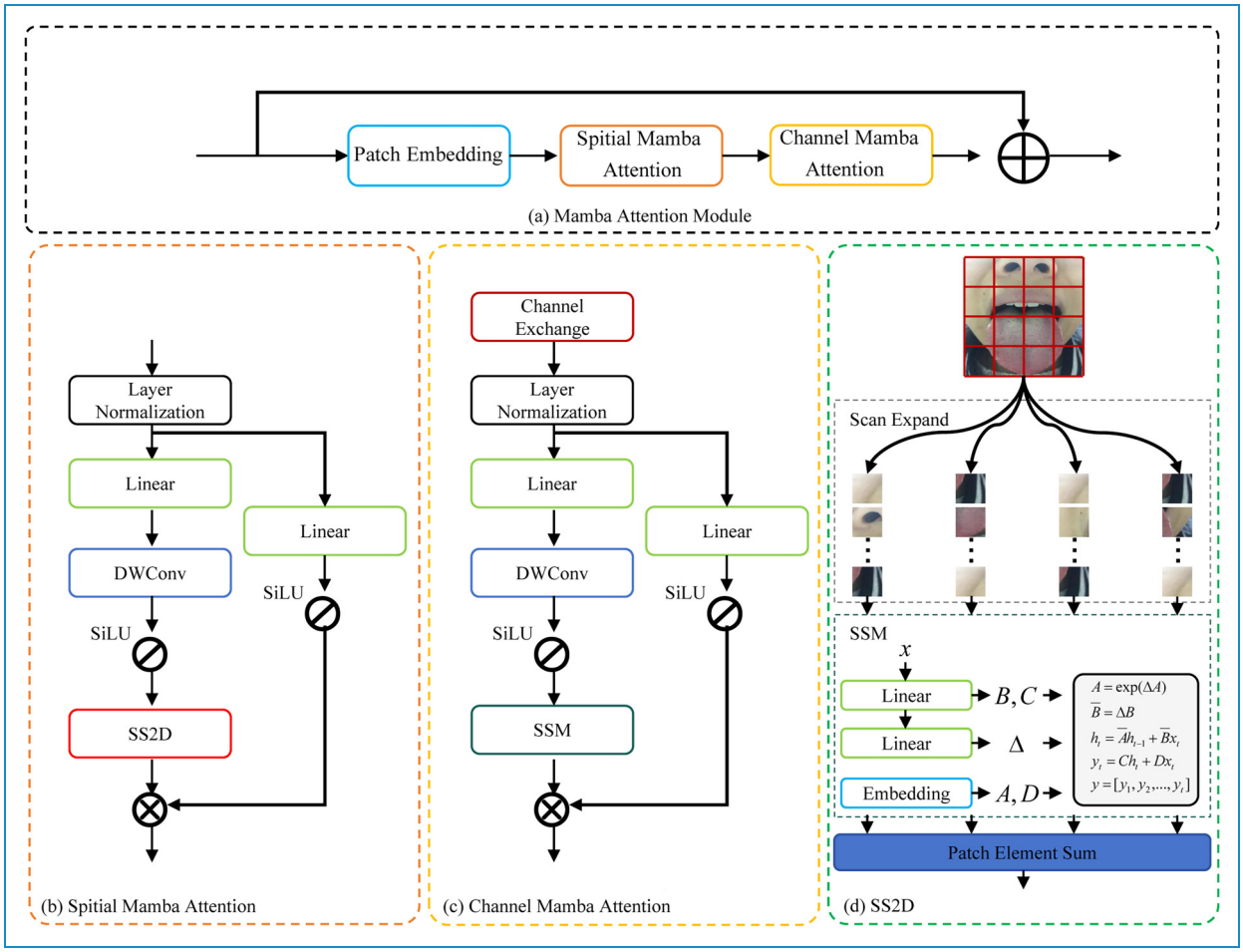

Figure 2. Structure of the Mamba attention module: (a) the overview of the Mamba attention module, (b) the structure of the spatial Mamba attention, (c) the structure of the channe Mamba attention, and (d) the structure of the SS2D and SSM. SS2D: 2D-selective-scan; SSM: state-space model.

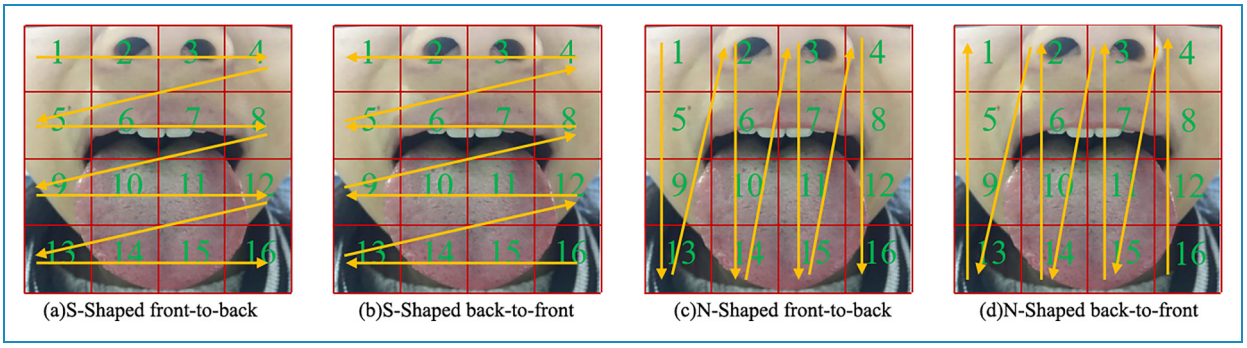

Figure 3. Four different patch sequences: (a) S-shaped front-to-back, (b) S-shaped back-to-front, (c) N-shaped front-to-back, and (d) N-shaped back-to-front.

from other positions and orientations, which in turn gives the model a global feel for the field. Mining quadruple connected domain space features of different patches by finding state space features for each of the four different sequences. Next, after the four sequential feature extraction is completed the results are re-correspondingly fused into a single image. Based on this, as shown in Figure 2(b), this study proposes a spatial Mamba attention similar to the spatial attention mechanism. It extracts image global spatial features through the advantage of Mamba mining long-range relations. It first normalizes a single sequence by layer normalization. Next, a GRU-like structure is used to achieve a similar effect to the attention mechanism. It consists of two main branches, the first of which aligns the sequence channels by a linear layer and a depth-separable convolution. After activation by SiLU, as shown in Figure 2(d) the sequence is divided into four sequences of different order for state space feature selection. Finally, it is multiplied by the linear features of the sequence extracted from the other branch. The above state space selection is specifically shown in equations (2) and (3). Where $\mathbf{X} = \{\mathbf{X_1}, \mathbf{X_2}, \mathbf{X_3}, \mathbf{X_4}\}$, $\mathbf{X_i} = \{\mathbf{x_i^1}, \mathbf{x_i^2}, \ldots, \mathbf{x_i^M}\}$. h stands for sequence state, j is the sequence order, and y is the output sequence. Similar to $\mathbf{X}$, $\mathbf{Y} = \{\mathbf{Y_1}, \mathbf{Y_2}, \mathbf{Y_3}, \mathbf{Y_4}\}$, $\mathbf{Y_i} = \{\mathbf{y_i^1}, \mathbf{y_i^2}, \ldots, \mathbf{y_i^M}\}$. $\bar{A}$, $\bar{B}$, $\bar{C}$, and $\bar{D}$ are the parameters of space-state model. In the research, we omit $\bar{D}$ to ease the calculation. $\bar{A}$ is set to $e^{\Delta A}$. $\bar{B}$ is set to ΔB. $\bar{C}$ is set to C. Where $B \in R^{B*(N*N)*H}$, $C \in R^{B*(N*N)*H}$, $\Delta \in R^{B*(N*N)*M}$ are derived from the input data. H is the hidden dim.

$$h_j = \bar{A}h_{j-1} + \bar{B}x_j \tag{2}$$

$$y_j = \bar{C}h_j + \bar{D}x_j \tag{3}$$

Multi-stage feature fusion module

The spatial features within the patch sequence were extracted through the aforementioned process. Similar to CBAM, we then apply a channel Mamba attention shown in Figure 2(c) for the channel feature extraction of the sequence after the spatial Mamba attention. First, we performed a dimensional transformation, swapping the channel with the patch. The input data shape of channel Mamba attention becomes $B * M * (N * N)$. The Mamba similar to the previous is followed. As channels are different from spatial, it is only required to focus on the features of the neighboring channels in the sequence. Therefore, we replace SS2D of Mamba with SSM. Finally, to further improve the computational efficiency, the residual element sum for the images before and after extracting the spatial and channel features.

The Mamba attention module efficiently extracts global features from the image in skip connection. In complex environments, capturing detailed information also requires accurate local detailed features. CNN can extract local features efficiently, the preservation of details diminishes as the number of layers increases. To address these problems, this paper proposes an early output method that uses the U-shaped network's structural properties to fuse multi-stage features, thereby enhancing detail acquisition ability. Its specific structure is illustrated in Figure 4. The input data of the multi-stage feature fusion module is the output from each stage. Since each stage outputs feature maps of different shapes and channels, they are first aligned. The alignment module is composed of an up-sampling module and a 1×1 convolution module. The size of each feature map is first changed by the up-sampling module into 512×512 by the inverse convolution operation. Because the feature map of the first stage is of size 512×512, it does not need to do upsampling. Then 1×1 convolution module is used to unify the channel size of each feature map. It is a convolution kernel with channel 1. In this way, we get five $512 \times 512 \times 1$ feature maps. Finally, the feature maps of each stage are fused to form the output feature map, which is further refined using a 1×1 convolution operation to produce the final output.

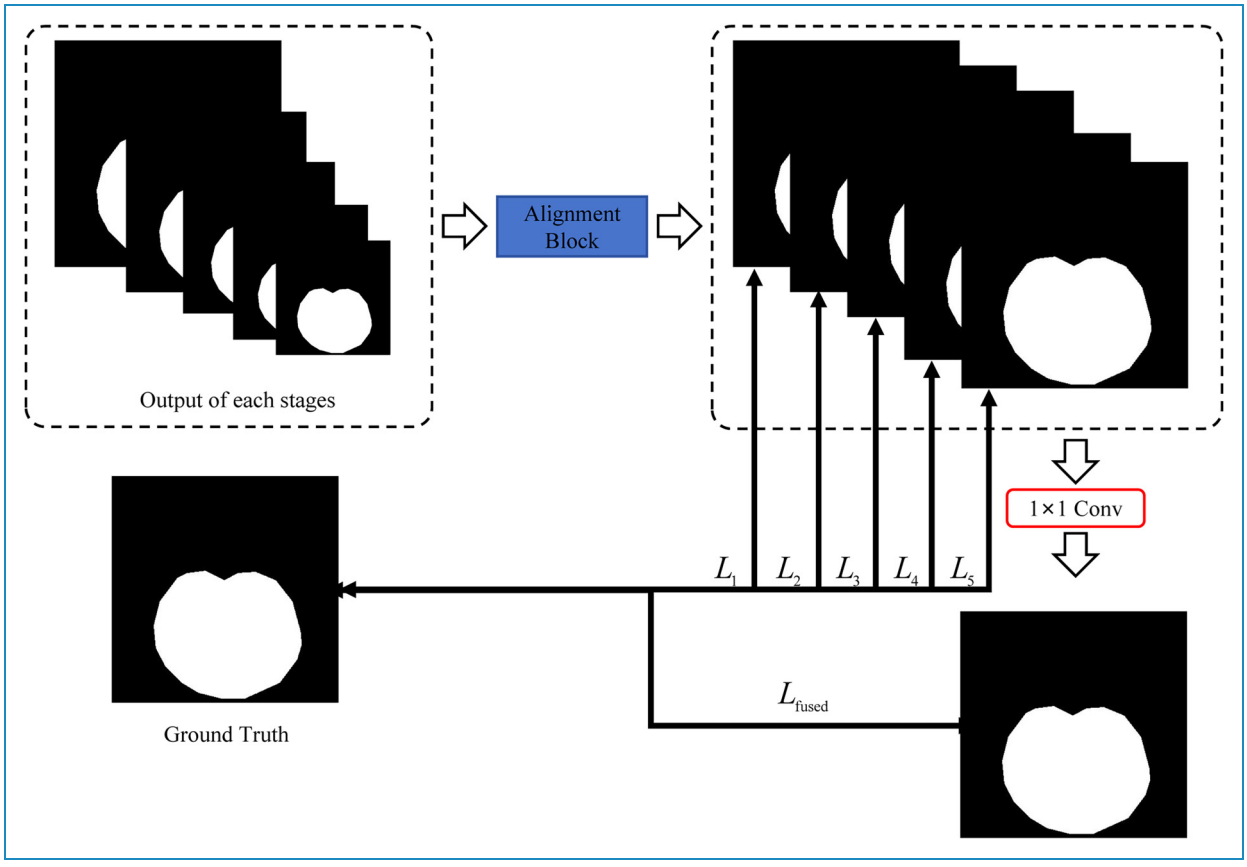

Figure 4. Structure of multi-stage feature fusion module.

Loss function

In this article, losses are categorized into stage loss and result loss. Stage loss is computed for each stage of the feature map, ensuring that the output of each stage retains as many detailed features as possible. The computation of this loss is specified by equations (4) and (5). S_i presents the aligned feature map of the i_{th} stage. *GroundTruth* denotes the target. The binary cross-entropy loss function is utilized, with equation (5) shows the calculation details. C indicating the class number.

$$L_i = BCE(S_i, GroundTruth) \tag{4}$$

$$BCE(y, \hat{y}) = -\sum_{i=1}^{C} (y_i log(\hat{y}_i) + (1 - y_i) log(1 - \hat{y}_i)) \tag{5}$$

Result loss, conversely, evaluates the discrepancy between the result feature map and the ground truth. It is calculated in equation (6). The total loss of the model is then determined by equation (7). *FusedMap* is the fused feature map. i represents the stage order.

$$L_{fused} = BCE(FusedMap, GroundTruth) \tag{6}$$

$$L_{total} = \sum_{i=1}^{5} L_i + L_{fused} \tag{7}$$

Results

In this section, we compare the tongue segmentation performance of the proposed model with existing model through metrics such as MIoU, precision, recall, and $F1$-score.

Datasets

This study utilized a publicly available dataset for its observational analysis. The tongue segmentation dataset includes images from two sources: TongueSAM.[50] They are 300 tongue images taken from Harbin Institute of Technology in 2014 and 1000 tongue images collected on the web in 2022. These datasets were merged to create a unified dataset of 1300 tongue images. The images come in two sizes: $576 \times 768 \times 3$ and $400 \times 400 \times 3$. To standardize the image sizes, we initially padded the image edges using the length of the longer edge as the target length. Subsequently, the images were resized to $512 \times 512 \times 3$.

Training details

The dataset was randomly divided into a training dataset and a test dataset, with a ratio of 4:1 for training to test data. All the models in the research are constructed by TensorFlow 2.6.0 and Keras 2.6.0. The training was

conducted on a single Nvidia 3090 16G GPU. The models were trained with a learning rate of 10^{-4} over 100 epochs, utilizing the Adam optimization algorithm. The batch size was set to 5.

Evaluation indicators

We used MIoU, precision, recall, and $F1$-score to assess segmentation performance. MIoU was the primary metric, while $F1$-score was the secondary evaluation indicator. The specific calculation methods for these metrics are shown in equations (8) to (11), where k in the equation presents the class number, TP denotes true positive, FN indicates false negative, and FP signifies false positive. Additionally, we compared their inference time durations.

$$MIoU = \frac{1}{k+1}\sum_{i=0}^{k}\frac{TP}{TP+FN+FP} \tag{8}$$

$$Precision = \frac{TP}{TP+FP} \tag{9}$$

$$Recall = \frac{TP}{TP+FN} \tag{10}$$

$$F1-Sscore = \frac{2*Precision*Recall}{Precision+Recall} \tag{11}$$

Segmentation performance comparison

In this section, we compared the proposed model from the "Materials and methods" section with other models using the previously mentioned metrics. The models were classified into four categories, the U-Net,[51] the models in which the attention guides the U-Net feature transfer using CBAM[16] or SENet,[52] the model for tongue segmentation in the complex environment such as OET Net[14] and RMP U-Net,[53] and the current best models for semantic segmentation of images such as Mobile Net,[54] U^2-Net,[55] and Swin Transformer.[56] The comparison results are shown in Table 1. The TUMamba demonstrates the best performance in both MIoU and $F1$-score metrics. Specifically, it improves MIoU by 1.17% and $F1$-score by 0.56%. In comparison with the semantic segmentation model Swin Transformer, which is currently the best performer in the tongue segmentation task, TUMamba improved the MIoU by 0.50% and the $F1$-score by 0.06% but reduced the inference time by 102.90 ms and model parameters by 73.96 MB. Although all of the attention-based models achieved high recall, their precision was low and showed the same over-segmentation as[13] mentioned.

We show the segmentation results of different models for images acquired in different environments since the dataset contains two types of images taken from the laboratory and complex environments. Figure 5 shows the results of different models for standard acquisition. In the standard setting, the different models have comparable effects on the segmentation of the tongue except for RMP U-Net. However, when examining the details of the segmentation results, particularly the center-upper region of the tongue, it becomes evident that models such as U-Net, SE U-Net, OET NET, Mobile Net, U^2-Net, and Swin Transformer do not achieve smooth segmentation in this area.

Table 1. Comparison of indicators.

Models	MIoU (%)	Precision (%)	Recall (%)	$F1$-score (%)	Inference time (ms)	Model prameters (MB)
U-Net	95.97	97.78	97.13	97.46	**24.95**	<u>7.36</u>
Attention U-Net	94.94	95.12	97.69	96.39	203.46	9.33
SE U-Net	87.63	83.70	**98.54**	90.52	201.17	9.37
OET NET	96.51	97.73	97.36	97.55	<u>40.80</u>	7.75
RMP U-Net	91.09	94.32	96.59	95.31	113.66	7.81
Mobile Net	96.64	97.84	98.08	<u>97.96</u>	148.53	**2.96**
U^2-Net	<u>96.76</u>	**98.06**	97.88	<u>97.96</u>	249.36	44.01
Swin Transformer	96.53	<u>98.01</u>	97.74	97.88	201.01	111.85
TUMamba (ours)	**97.14**	97.85	<u>98.19</u>	**98.02**	146.46	37.87

MIoU: mean intersection over union.
Bold is the best result, underline is the second best.

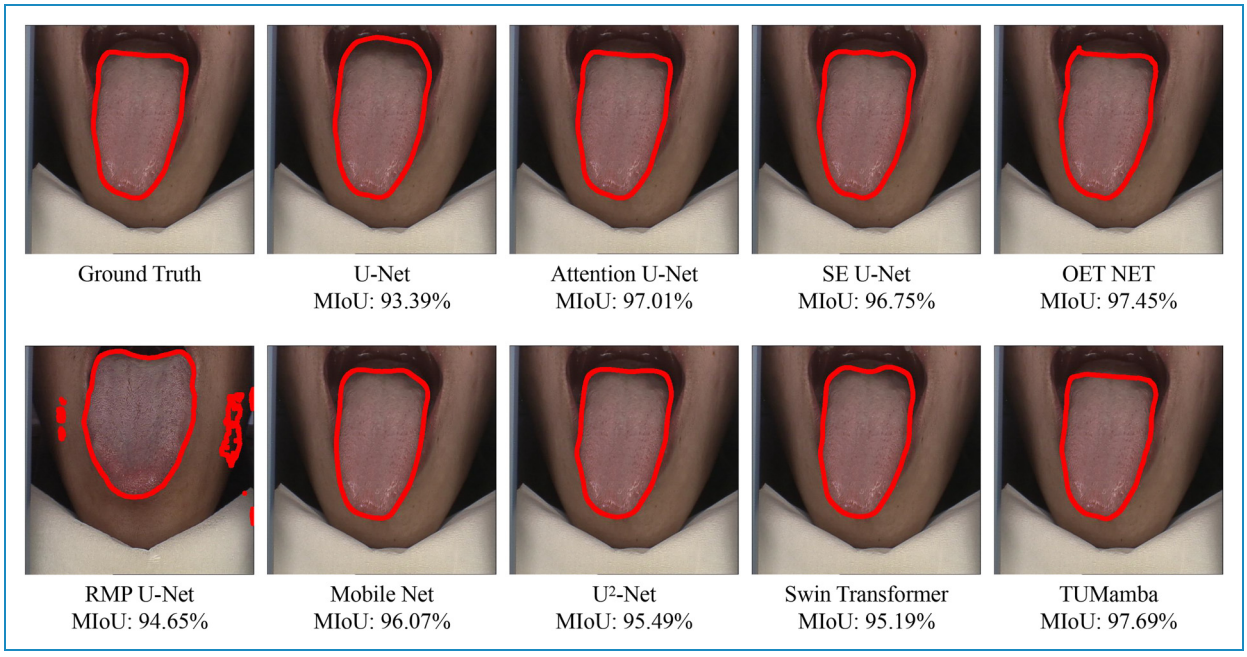

Figure 5. Results of different models for tongue segmentation for standard acquisition.

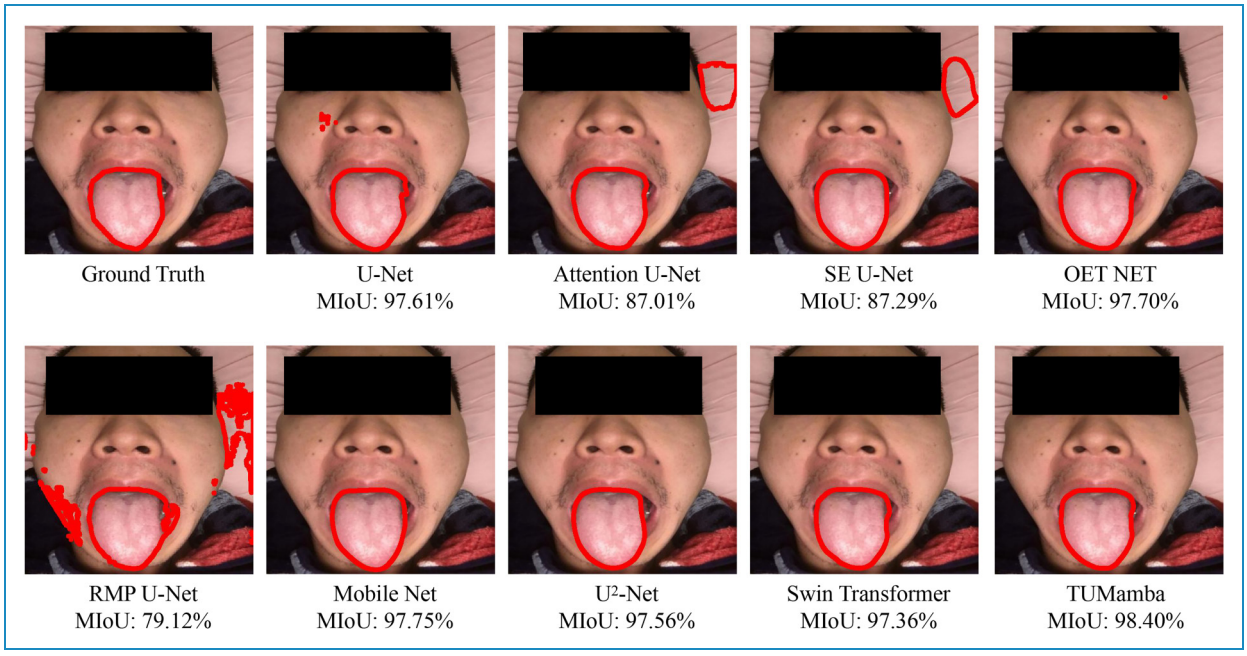

Figure 6. Results of different models for tongue segmentation for complex environment acquisition.

Figure 6 shows the results of different models for complex environment acquisition. The U-Net exhibits over-segmentation caused by the background and clothing. Interestingly, the Attention model not only failed to reduce over-segmentation but exacerbated it. Regarding detail extraction, such as curved depression on the right side of the tongue, both SE U-Net and Mobile Net models were unable to effectively capture this feature.

Ablation experiment

We performed ablation experiments using U-Net as the backbone to evaluate the impact of adding and removing modules

Table 2. Results of ablation experiment.

Models	MIoU (%)	Precision (%)	Recall (%)	*F*1-score (%)
U-Net	95.97	97.78	97.13	97.46
U-Net + Mamba attention module	96.90	97.81	97.95	97.88
U-Net + multi-stage feature fusion module	96.43	97.73	97.45	97.59
U-Net + Mamba attention module + multi-stage feature fusion module	**97.14**	**97.85**	**98.19**	**98.02**

MIoU: mean intersection over union.
Bold is the best result.

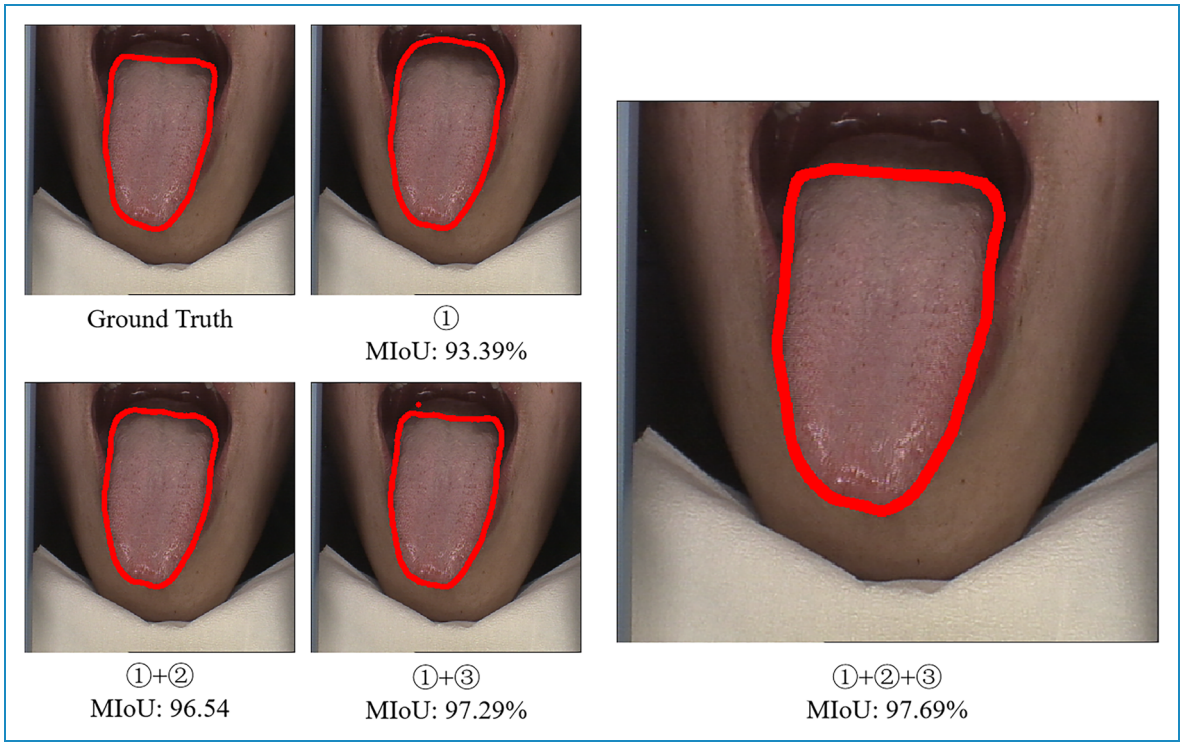

Figure 7. Results of different module additions and subtractions for tongue segmentation for laboratory environment acquisition. 1: U-Net, 2: Mamba attention module, and 3: multi-stage feature fusion module.

on tongue segmentation performance. The results, detailed in Table 2. For single-module additions, the Mamba attention module notably improved the segmentation performance, increasing the MIoU by 0.46% compared to the backbone U-Net model. As additional modules were stacked, the model's efficacy continued to improve. The TUMamba model, which combines the Mamba attention module and the multi-stage feature fusion module with U-Net, achieved the best segmentation results, improving the MIoU by 1.17% over the backbone model.

Figure 7 illustrates the effect of different module additions and subtractions on tongue segmentation under laboratory conditions. The results indicate that these modifications affect tongue segmentation outcomes in a controlled environment. While multi-feature map fusion may introduce some ambient noise, the Mamba attention module effectively reduces noise interference.

Figure 8 displays the results of the ablation experiment in a complex interference environment. Although multi-feature map fusion does not completely address environmental noise interference, it enhances the smoothness of the segmentation results. The Mamba attention module helps the model mitigate environmental noise interference effectively.

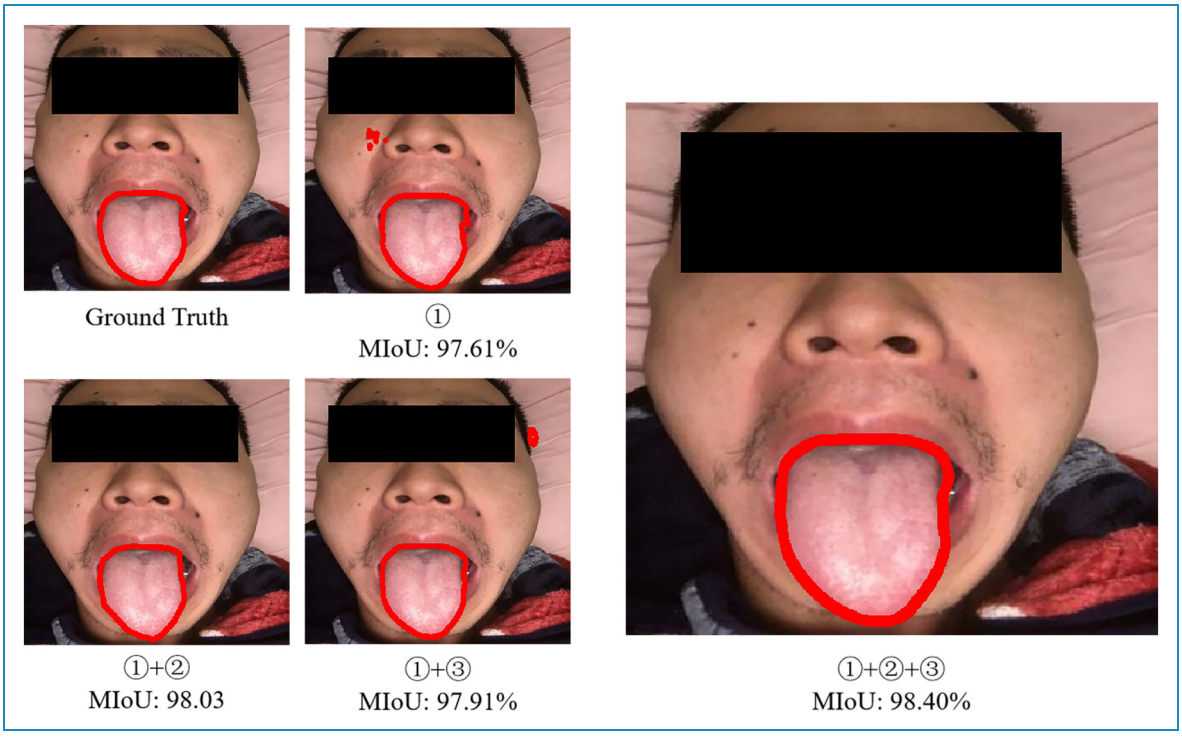

Figure 8. Results of different module additions and subtractions for tongue segmentation for complex environment acquisition. 1: U-Net, 2: Mamba attention module, and 3: multi-stage feature fusion module.

Table 3. Results of different structures.

Models	MIoU (%)	Precision (%)	Recall (%)	*F*1-score (%)
Channel∪ spatial	96.99	97.45	97.78	97.61
Channel+spatial	97.05	97.75	97.98	97.86
Spatial+channel	**97.14**	**97.85**	**98.19**	**98.02**

Bold is the best result.

Discussion

U-Net is widely recognized for its efficiency in image processing tasks, due to its effective capture of local features and its distinctive U-shaped structure, which integrates both shallow and deep information. Despite these advantages, U-Net struggles with capturing global features and is prone to transferring noise into semantic information through its skip connections. This limitation becomes evident in complex environments, as shown in Figure 6, where U-Net exhibits significant over-segmentation due to the presence of tongue-like objects. Traditionally, attention mechanisms have been used to address such issues. However, our experiments corroborated the findings of,[14] which suggests that incorporating attention into U-Net paradoxically, degrade performance. Analysis of the metrics Table 1 suggests that this poor performance is similar to over-fitting. While attention mechanism enhance the model's recall, they concurrently reduce precision, resulting in segmentation outputs that include more of the target but also increased background noise. This problem stems from a mismatch between the sample size of the image dataset and variability in background features, as the training samples are adequate for handling the highly variable backgrounds. To address these challenges, we employed the recently introduced Mamba attention mechanism. Spatial Mamba attention enhances the model's ability to manage background noise by processing four different order sequences and dividing images into smaller patches to extract global spatial information. Furthermore, to improve global channel feature extraction, we integrated

channel Mamba attention following spatial Mamba attention, inspired by the CBAM framework.[16] As shown in Table 2, the Mamba attention module significantly improves both precision and recall, effectively resolving the over-segmentation issue. In conclusion, the proposed model addresses both over-segmentation and under-segmentation issues in tongue segmentation tasks within complex environments, achieving accurate separation of the tongue from tongue-like objects. Compared to current popular mobile segmentation models listed in Table 1, our model's inference time is comparable to that of Mobile Net, indicating its potential for deployment on mobile devices for rapid and precise segmentation.

Our model's inference time is comparable to that of Mobile Net, indicating its potential for deployment on mobile devices for rapid and precise segmentation. Figure 6, shows that while each model performs similarly in standard environments, there is still a gap in edge detail handling. To capture detailed information from each feature map, we fused the feature maps at each stage and applied robust supervision to the tongue image detail segmentation by calculating the loss for each stage's feature maps. As shown in Table 2, this method significantly enhances the model's segmentation performance, with Figure 5 illustrating that it effectively extracts more detailed features of the tongue edge.

In this study, we explored two common methods for fusing spatial and channel attention: serial and parallel connections. Serial connections, as utilized by Sanghyun et al.[16] and parallel connections, as described by Fu et al.,[35] each offer distinct advantages. Based on the experimental results presented in Table 3, we chose a CBAM-like approach with serially connected attention mechanisms. While parallel extraction can improve feature extraction efficiency, it is less effective compared to serial connection, as it may fail to adequately filter out noise. Moreover, channel feature extraction first, as opposed to spatial feature extraction, can be less effective with Mamba due to potential loss of spatial features during sequence processing. Thus, a CBAM-like feature extraction architecture was adopted in our research.

Conclusion

In this article, we introduced a novel tongue segmentation model, leveraging both Mamba and U-Net architectures. Our approach incorporated a Mamba attention module to extract global spatial and channel features effectively. To maintain fine details and local information, we enforced strong supervision at each stage of the model, with the final results being a fusion of outputs from all stages. Our method excelled particularly in tongue segmentation tasks. However, it's worth noting that model based on Mamba struggles with extracting precise location information. Although our method addressed this issue to some extent by employing four different sequences for location feature extraction, there remained a disparity compared to traditional convolution kernels. Therefore, our future main work will focus on Mamba for 2D image spatial information extraction. In this way, accurate extraction of images based on Mamba can be achieved. Then, combined with Mamba's advantage of text feature extraction, we will construct a multi-modal macro-model of TCM.

Acknowledgements: We thank Cao et al.[50] for the open source tongue segmentation database. The public dataset is acquired at https://github.com/cshan-github/TongueSAM.

Contributorship: Fan Jiang: conceptualization; Fan Jiang: methodology; Fan Jiang: software; Fan Jiang and Yanmei Zhong: validation; Fan Jiang: writing–original draft preparation; Yanmei Zhong and Simin Yang: writing–review and editing; Simin Yang: supervision.

Consent to participate: Not applicable. The data used in this article comes from publicly available datasets, so informed patient consent is not required for this research.

Declaration of conflicting interests: The author(s) declared no potential conflicts of interest with respect to the research, authorship, or publication of this article.

Ethical approval: The data used in this article comes from publicly available datasets, so there are no ethical conflicts.

Funding: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported in part by the Chengdu medical research project (Fund Number: 2020176).

Guarantor: FJ.

ORCID iD: Simin Yang https://orcid.org/0009-0005-1245-3116

References

1. Felix C. TCM: made in China. *Nature* 2011; 480: S82–S83.
2. David C. Why Chinese medicine is heading for clinics around the world. *Nature* 2018; 561: 448–450.
3. Meng XL, Liu XQ, Tan JY, et al. From Xiaoke to diabetes mellitus: a review of the research progress in traditional Chinese medicine for diabetes mellitus treatment. *Chin Med* 2023; 18: 1–22.
4. Chen LS, Wei SZ, He Y, et al. Treatment of chronic gastritis with traditional Chinese medicine: pharmacological activities and mechanisms. *Pharmaceuticals* 2023; 16: 1–16.
5. Shi MM, Piao JH, Xu XL, et al. Chinese medicines with sedative-hypnotic effects and their active components. *Sleep Med Rev* 2016; 29: 108–118.

6. Wang XZ, Zhang B, Guo ZH, et al. Facial image medical analysis system using quantitative chromatic feature. *Expert Syst Appl* 2013; 40: 3738–3746.
7. Li JW, Zhang ZD, Zhu XL, et al. Automatic classification framework of tongue feature based on convolutional neural networks. *Micromachines* 2022; 13: 1–12.
8. Li QF, Chen YH, Pang YJ, et al. An AAM-based identification method for ear acupoint area. *Biomimetics* 2023; 8: 1–16.
9. Zhang Q, Zhou JH and Zhang B. Computational traditional Chinese medicine diagnosis: a literature survey. *Comput Biol Med* 2021; 133: 104358.
10. Ma C, Yu M, Chen FK, et al. An efficient and portable LED multispectral imaging system and its application to human tongue detection. *Appl Sci* 2022; 12: 3552.
11. Vibha B and Prashant P. Challenges and solutions in automated tongue diagnosis techniques: a review. *Crit Rev Biomed Eng* 2022; 50: 47–63.
12. Ozan O, Jo S, Loic LF, et al. Attention U-Net: learning where to look for the pancreas. *Arxiv*, 2018.
13. Shi PF, Xu XW, Fan XN, et al. LL-UNet++:UNet++ based nested skip connections network for low-light image enhancement. *IEEE Trans Comput Imaging* 2024; 10: 510–521.
14. Huang ZH, Miao JQ, Song HB, et al. A novel tongue segmentation method based on improved U-Net. *Neurocomputing* 2022; 500: 73–89.
15. Gu A and Dao T. Mamba: linear-time sequence modeling with selective state spaces. *Arxiv*, 2023.
16. Sanghyun W, Jongchan P, Lee JY, et al. CBAM: convolutional block attention module. In: *European conference on computer vision*, 2018.
17. Du JQ, Lu YS, Zhu MF, et al. A novel algorithm of color tongue image segmentation based on HSI. In: *International conference on biomedical engineering and informatics*, 2008.
18. Chen L, Wang DY, Liu YQ, et al. A novel automatic tongue image segmentation algorithm: color enhancement method based on L*a*b* color space. In: *IEEE international conference on bioinformatics and biomedicine*, 2015.
19. Huang YS, Zhang Q and Huang ZP. Tongue image segmentation based on the sub-block region growing algorithm. In: *International conference on information science and control engineering*, 2016.
20. Saparudin E and Muhammad F. Tongue segmentation using active contour model. In: *IAES international conference on electrical engineering, computer science and informatics*, 2017.
21. Gu HY, Yang ZC and Chen H. Automatic tongue image segmentation based on thresholding and an improved level set model. In: *IEEE Asia Pacific conference on circuits and systems*, 2020.
22. Zhou JH, Zhang Q, Zhang B, et al. tongueNet: a precise and fast Tongue segmentation system using U-Net with a morphological processing layer. *Appl Sci* 2019; 9: 3128.
23. Feng L, Huang ZH, Zhong YM, et al. Research and application of tongue and face diagnosis based on deep learning. *Digital Health* 2022; 8: 20552076221124436.
24. Xu Q, Zeng Y, Tang WJ, et al. Multi-task joint learning model for segmenting and classifying tongue images using a deep neural network. *IEEE J Biomed Health Inform* 2020; 24: 2481–2489.
25. Huang ZH, Huang WC, Wu HC, et al. TongueMobile: automated tongue segmentation and diagnosis on smartphones. *Neural Comput Appl* 2023; 35: 21259–21274.
26. Song HB, Huang ZH, Feng L, et al. RAFF-Net: an improved tongue segmentation algorithm based on residual attention network and multiscale feature fusion. *Digital Health* 2022; 8: 20552076221136362.
27. Ruan QS, Wu QF, Yao JF, et al. An efficient tongue segmentation model based on U-Net framework. *Int J Pattern Recognit Artif Intell* 2021; 35: 2154035.
28. Zhan GL, Qian K, Chen WY, et al. EAswin-unet: segmenting CT images of COVID-19 with edge-fusion attention. *Biomed Signal Process Control* 2023; 89: 105759.
29. Sadeghibakhi M, Hamidreza P, Hamidreza M, et al. Multiple sclerosis lesions segmentation using attention-based CNNs in FLAIR images. *IEEE J Transl Eng Health Med* 2022; 10: 1800411.
30. Liu HL, Feng Y, Xu H, et al. MEA-Net: multilayer edge attention network for medical image segmentation. *Sci Rep* 2022; 12: 7868.
31. Li LH, Zhang WJ, Zhang XY, et al. Semi-supervised remote sensing image semantic segmentation method based on deep learning. *Electronics* 2023; 12: 348.
32. Lu JF, Shi MT, Song CH, et al. Classifier-guided multi-style tile image generation method. *J King Saud Univ Comput Inf Sci* 2024; 36: 101899.
33. Zhu M, Wang JC, Wang A, et al. Multi-fusion approach for wood microscopic images identification based on deep transfer learning. *Appl Sci* 2021; 11: 7639.
34. Lu JF, Ren HP, Shi MT, et al. A novel hybridoma cell segmentation method based on multi-scale feature fusion and dual attention network. *Electronics* 2023; 12: 979.
35. Fu J, Liu J, Tian HJ, et al. DANet: dual attention network for scene segmentation. In: *IEEE/CVF conference on computer vision and pattrern recognition*, 2019, pp.3146–3154.
36. Li DM, Yin SY, Lei Y, et al. Segmentation of white blood cells based on CBAM-DC-UNet. *IEEE Access* 2023; 11: 1074–1082.
37. Shah HA and Kang JM. An optimized multi-organ cancer cells segmentation for histopathological images based on CBAM-residual U-Net. *IEEE Access* 2023; 11: 111608–111621.
38. Wang YQ, Ma JW, Sergey A, et al. DPA-UNet rectal cancer image segmentation based on visual attention. *Concurrency Comput Pract Exper* 2023; 35: e7670.
39. Huang SQ, Li JN, Xiao YZ, et al. RTNet: relation transformer network for diabetic retinopathy multi-lesion segmentation. *IEEE Trans Med Imaging* 2022; 41: 1596–1607.
40. Zhao L, Tan GH, Pu B, et al. TransFSM: fetal anatomy segmentation and biometric measurement in ultrasound images using a hybrid transformer. *IEEE J Biomed Health Inform* 2024; 28: 285–296.
41. Yang JY, Jiao LC, Shang RH, et al. EPT-Net: edge perception transformer for 3D medical image segmentation. *IEEE Trans Med Imaging* 2023; 42: 3229–3243.
42. Fu BK, Peng YS, He JJ, et al. HmsU-Net: a hybrid multi-scale U-net based on a CNN and transformer for medical image segmentation. *Comput Biol Med* 2024; 170: 108013.
43. Chen ZY, Chen SY and Hu FJ. CTA-UNet: CNN-transformer architecture UNet for dental CBCT images segmentation. *Phys Med Biol* 2023; 68: 175042.
44. Yu A, Shi WL, Ji B, et al. MS-TCNet: an effective transformer-CNN combined network using multi-scale feature learning for 3D medical image segmentation. *Comput Biol Med* 2024; 170: 108057.

45. Gu A, Karan G and Christopher R. Efficiently modeling long sequences with structured state spaces. *Arxiv*, 2021.
46. Jimmy THS, Andrew W and Scott WL. Simplified state space layers for sequence modeling. *Arxiv*, 2022.
47. Harsh M, Ankit G, Ashok C, et al. Simplified state space layers for sequence modeling. *Arxiv*, 2022.
48. Zhu LH, Liao BC, Zhang Q, et al. Vision Mamba: efficient visual representation learning with bidirectional state space model. *Arxiv*, 2024.
49. Liu Y, Tian YJ, Zhao YZ, et al. VMamba: visual state space model. *Arxiv*, 2024.
50. Cao S, Ruan Q and Ma L. TongueSAM: an universal tongue segmentation model based on SAM with zero-shot. *Arxiv*, 2023.
51. Olaf R, Philipp F and Thomas B. U-Net: convolutional networks for biomedical image segmentation. *Med Image Comput Comput Assist Interv* 2015; 9351: 234–241.
52. Abhijit GR, Nassir N and Christian W. Concurrent spatial and channel squeeze & excitation in fully convolutional networks. In: *International conference on medical image computing and computer*, 2018.
53. Worapan K, Panyanuch B, Sarattha K, et al. Encoder-decoder network with RMP for tongue segmentation. *Med Biol Eng Comput* 2023; 61: 1193–1207.
54. Howard AG, Zhu M, Chen B, et al. MobileNets: efficient convolutional neural networks for mobile vision applications. *Arxiv*, 2017.
55. Qin XB, Zhang ZC, Huang CY, et al. U2-Net: going deeper with nested U-structure for salient object detection. *Pattern Recognit* 2020; 106: 107404.
56. Liu Z, Lin YT, Cao Y, et al. Swin transformer: hierarchical vision transformer using shifted windows. In: *IEEE/CVF international conference on computer vision*, 2021, pp.10012–10022.