AI Medicine
https://www.sciltp.com/journals/aim

Article

CNN-Based Tongue Image Segmentation in Traditional Chinese Medicine

Dechao Xu [1] and Dingcheng Tian [1,2,*]

[1] Research Institute for Medical and Biological Engineering, Ningbo University, Ningbo 315211, China
[2] College of Medicine and Biological Information Engineering, Northeastern University, Shenyang 110016, China
* Correspondence: 2310520@stu.neu.edu.cn

How To Cite: Xu, D.; Tian, D. CNN-Based Tongue Image Segmentation in Traditional Chinese Medicine. *AI Medicine* **2025**, *2*(2), 4. https://doi.org/10.53941/aim.2025.100004

Received: 28 April 2025
Revised: 15 July 2025
Accepted: 1 August 2025
Published: 12 August 2025

Abstract: Tongue image segmentation is a key component of the intelligent diagnosis in Traditional Chinese Medicine (TCM), and its accuracy directly affects the subsequent classification and diagnostic results. However, current research faces challenges due to data scarcity and limited model adaptability. The lack of publicly available tongue image datasets restricts the model's generalization ability, while traditional algorithms are highly sensitive to lighting and posture changes. Moreover, general deep learning models have not been optimized for the characteristics of blurry tongue edges and low contrast background, which often leads to loss of details or over-segmentation. To address these issues, this study has constructed a high-quality dataset containing 1405 tongue images and proposes a lightweight tongue image segmentation network (TSNet). TSNet employs depthwise separable convolutions to reduce computational cost, introduces a Parallel Atrous Spatial Pyramid Pooling (PASPP) module to extract multi-scale features and handle blurred boundaries, and incorporates a Boundary Adjustment (BA) module to enhance edge segmentation accuracy. Experimental results show that TSNet achieves a mean Intersection over Union (mIoU) of 97.20% while using fewer parameters than mainstream models. While preserving tongue details, it effectively reduces the number of parameters, providing an efficient solution for tongue image segmentation in TCM.

Keywords: deep learning; traditional Chinese medicine; tongue image segmentation; convolutional neural network

1. Introduction

In the long history of clinical diagnosis in Traditional Chinese Medicine (TCM), four basic diagnostic methods, inspection, auscultation, inquiry, and palpation, have been established [1]. Among them, tongue diagnosis, as the most direct inspection method, has long been used to assess the body's health status. The shape, color, and coating of the tongue can reflect the body's Qi, blood, Yin-Yang balance, and the function of internal organs [2]. However, traditional tongue diagnosis relies heavily on the practitioner's experience, which is subjective and lacks quantitative description, limiting its broader clinical application.

The deep integration of artificial intelligence and medical imaging technology provides a new path to address the above challenges. By using computer vision to quantify tongue features and applying deep learning to establish tongue image-syndrome correlation models, TCM tongue diagnosis is transitioning from an experience-based approach to a data-driven paradigm [3]. However, during the acquisition of tongue images, due to limitations of the capture equipment, the images often include facial background elements such as lips and teeth. Therefore, it is necessary to segment the tongue to prevent the background from affecting the analysis. As a fundamental step in intelligent tongue diagnosis, the accuracy of tongue image segmentation directly impacts the reliability of subsequent analysis.

Traditional tongue image segmentation methods, such as thresholding, edge detection, or active contour models, typically rely on manually designed features and rules [4]. Although some success has been achieved in specific scenarios, these methods have limited general applicability and are highly sensitive to lighting and tongue posture changes. Therefore, overcoming the limitations of these traditional methods and improving the accuracy and robustness of tongue image segmentation has become a challenging task in the research field.

In recent years, deep learning technologies have made significant breakthroughs in the medical field [5–9]. In particular, Convolutional Neural Networks (CNNs) have also been gradually applied to tongue image segmentation. Huang et al. proposed an automatic tongue image segmentation method based on an enhanced fully convolutional network. This method uses a deep residual network as the encoder, receptive field modules and a Feature Pyramid Network (FPN) decoder, effectively capturing global contextual information and fusing multi-scale feature maps to recover the tongue contour. Quantitative evaluation of the SIPL-tongue dataset shows that this method outperforms four other deep learning segmentation methods regarding average Hausdorff distance, Dice similarity coefficient, accuracy, and sensitivity, demonstrating its potential in automated tongue diagnosis [10]. Yao et al. proposed a tongue image segmentation method combining an improved U-Net and edge optimization post-processing. Through data augmentation, precise network design, and edge optimization, experiments show that this method performs excellently on multiple datasets and improves the segmentation results of classic neural networks [11]. Jia et al. proposed the QA-TSN model, which addresses the small sample problem through a Tongue Style Transfer Generation Network (T-STGN) and uses an improved partial convolution to accelerate real-time segmentation. The proposed tongue segmentation loss function (TSL) effectively smooths the tongue boundary. Experimental results show that QA-TSN outperforms other methods in segmentation accuracy and frame rate [12]. Huang et al. proposed the PriTongueNet model, which improves tongue image segmentation accuracy using an attention-guided module and geometric prior loss. Experiments show that the model performs excellently on two datasets, with inference time only one-third that of other models. The geometric prior loss significantly improves the segmentation performance and can be applied to different network architectures [13]. Although existing deep learning models, such as U-Net and DeepLab, have demonstrated outstanding performance in various medical image segmentation tasks, the unique challenges posed by TCM tongue images—such as low contrast between the tongue area and the background, and unclear tongue boundaries—mean that general-purpose deep learning models have not been optimized for these characteristics, resulting in issues such as detail loss or over-segmentation.

This study proposes a tongue image segmentation network (TSNet) based on CNN to address these issues. The network integrates depthwise separable convolutions, parallel atrous spatial pyramid pooling modules [14], and boundary optimization mechanisms, aiming to improve the accuracy and efficiency of tongue image segmentation. By constructing a high-quality dataset containing 1405 tongue images, this study provides an effective solution for tongue image segmentation in TCM. Experimental results show that TSNet preserves tongue details and reduces the number of parameters. Despite the challenges posed by complex backgrounds and diverse tongue images, TSNet achieves a mean Intersection over Union (mIoU) of 97.20%, demonstrating its potential in the intelligent automation of TCM tongue diagnosis. The main contributions of this paper are summarized as follows.

(1) Constructed a large-scale, high-quality dataset in the field of TCM tongue diagnosis, covering multi-source and diverse tongue images, which addresses the lack of publicly available data and provides a standardized foundation for future research.

(2) Proposed an improved PASPP module with optimized atrous convolution settings to effectively enhance multi-scale feature extraction, specifically targeting challenges such as blurred tongue boundaries and strong background interference.

(3) Designed a Boundary Adjustment (BA) module that significantly improves the model's ability to extract tongue edge contours, making it particularly suitable for fine-grained segmentation of complex tongue shapes in TCM applications.

(4) Developed an end-to-end TSNet architecture that achieves high segmentation accuracy (97.20% mIoU) and high pixel accuracy (98.52% MPA (mean pixel accuracy)), while significantly reducing model complexity with fewer parameters (48.74 M). Compared to other models, TSNet outperforms in segmentation accuracy, improving by 0.48% over FusionNet (FusionNet mIoU: 96.72%) and 2.49% over GCN (GCN mIoU: 94.71%). At the same time, TSNet has significantly fewer parameters than FusionNet (81.68 M) and GCN (58.14 M), demonstrating its potential in computational efficiency and practical applications.

The rest of this paper is organized as follows. Section 2 introduces the proposed model, Section 3 presents the experimental data and details of the experiments and analyzes the results, and finally, Section 4 provides the discussion and conclusion.

2. Method

The TSN model consists of three parts: feature encoding, contextual awareness, and feature decoding, as shown in Figure 1. After inputting the tongue image, the model performs feature extraction through multiple layers of convolution, batch normalization, and activation functions. To reduce parameters, depthwise separable convolutions and max pooling operations are used. Next, adaptive pooling and the ASPP module are applied to enhance the image's localization and recovery capabilities. The feature decoding part gradually restores the feature map details through operations like transposed convolution, pooling indices, and boundary adjustment (BA) while adjusting the channel numbers and sizes of the feature maps. The final output is the optimized result. The overall architecture improves feature extraction accuracy while effectively enhancing boundary detail recovery.

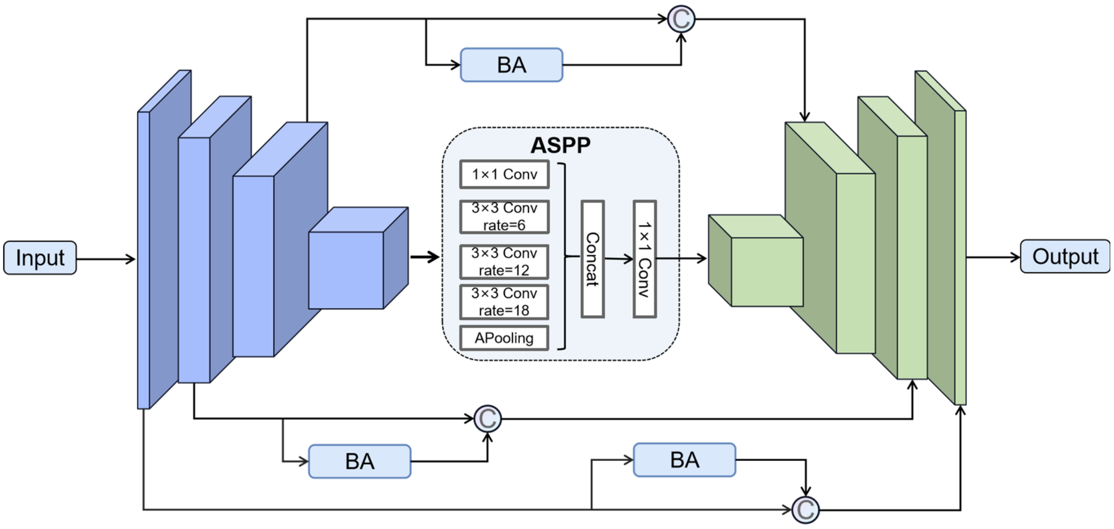

Figure 1. The overall architecture of the TSNet.

2.1. Feature Encoding Module

The feature encoding part consists of four encoder blocks, each comprising two convolutional layers. At the output of each convolutional layer, activation functions, and batch normalization mechanisms are added to enhance the network's non-linear expressiveness and training stability. To effectively reduce the model's parameter count and lower the risk of overfitting, some traditional convolutional layers are replaced with depthwise separable convolutions. Depthwise separable convolutions decompose the convolution operation, significantly reducing computational complexity while maintaining model performance.

Depthwise separable convolution can be divided into two main operational processes. The first is the depthwise convolution operation, during which each input channel is convolved with an independent convolution kernel, generating a set of output channels. Since each output channel only depends on a single input channel, this operation efficiently processes each channel, reducing computational complexity. Next is the pointwise convolution operation, where a 1 × 1 convolution kernel processes all pixels. Each output pixel only depends on the pixel value at the corresponding position in the input image, thus enabling cross-channel feature integration. This depthwise separable convolution structure effectively compresses the computation load and has been proven effective in various tasks, making it an essential technique for building efficient convolutional neural networks.

2.2. Contextual Awareness Module

The contextual awareness module consists of the Atrous Spatial Pyramid Pooling (ASPP) module, which aims to expand the receptive field and capture global information. This module is located between the feature encoding extraction and decoding modules, serving as a transition between encoding and decoding. This study designed two architectures to evaluate and compare the performance of ASPP. Parallel Atrous Spatial Pyramid Pooling (PASPP) and Series Atrous Spatial Pyramid Pooling (SASPP) were trained and evaluated for performance. Figure 2 shows the structures of SASPP and PASPP.

The PASPP module comprises four dilated convolution modules and one global pooling module. Each dilated convolution module includes a dilated convolution operation, batch normalization layer, and activation function. To avoid the grid effect caused by non-orthogonal convolutions, the dilation rates of each dilated convolution module are set to 1, 6, 12, and 18. The input data is first pooled through the global pooling operation and then combined with the outputs of the four dilated convolution modules through upsampling, ultimately producing the network's output. The SASPP module, on the other hand, consists of four dilated convolution modules. After the input data is processed by the four dilated convolution module (DCM), the output is further processed through a batch normalization (BN) layer to ensure the stability and robustness of the output features. The ASPP architecture effectively enhances the model's ability to perceive information at different scales, allowing it to capture essential features in various contextual environments, thereby improving its overall performance.

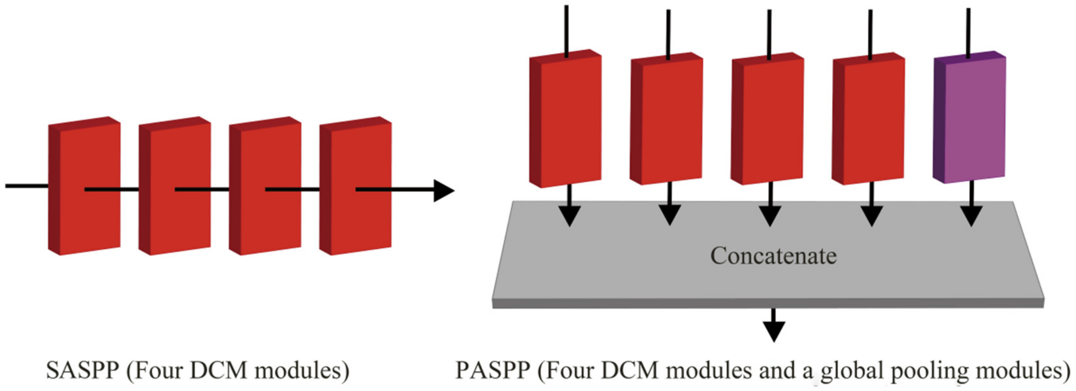

Figure 2. Structure of SASPP and PASPP.

2.3. Feature Decoding Module

The feature decoding module aims to restore the high-level semantic features and spatial information extracted by the feature encoding module. This module consists of four decoding blocks, including an upsampling module, a boundary adjustment module, and a 1 × 1 convolutional layer. The output of each decoding block is fused with the production of the feature encoding module, gradually restoring the image's detailed information. The Pooling Index (PI) [15] mechanism is introduced to further improve the tongue's localization and reconstruction. Through this mechanism, the feature encoding and decoding modules can more effectively fuse global and local information, thereby enhancing the model's localization accuracy and reconstruction performance.

BA module is added to the decoding module to enhance the boundary segmentation performance. The structure of the BA module is shown in Figure 3. The BA module has two parallel paths. The first path consists of a 1 × 1 convolutional module, while the second path comprises standard convolution and 1 × 1 convolution layers. After processing through batch normalization and activation function layers, the outputs of both paths are summed, resulting in the final output of the module. This structure effectively refines boundary information, optimizing the output of the decoding module and improving the model's boundary accuracy in segmentation tasks. Through this design, the feature decoding module not only restores the spatial structure of the image but also enhances the model's performance in fine boundary segmentation and accurate localization, thereby boosting overall performance.

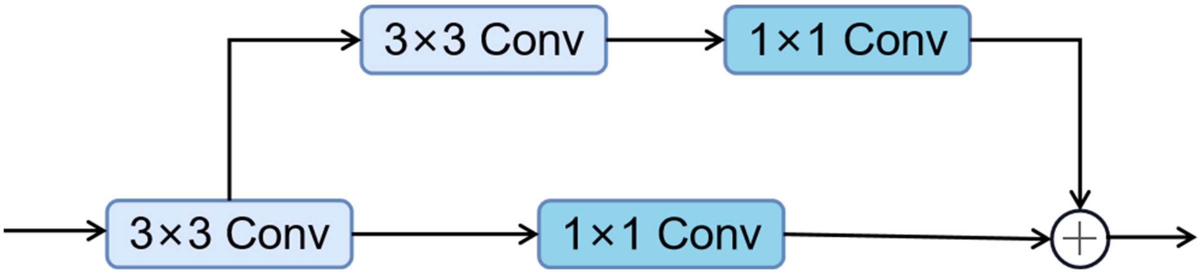

Figure 3. Structure of boundary adjustment module.

3. Experiments and Results

3.1. Dataset

This study systematically constructs a tongue image dataset to address the current lack of publicly available datasets in traditional Chinese medicine tongue diagnosis. Through cross-institutional collaboration and a multi-source data collection strategy (including offline collection, research project accumulation, and online public data screening), 1405 standardized tongue images were integrated. The details of the dataset are shown in Table 1. The tongue contour segmentation annotations were completed using the Labelme software labelme 5.1.1, establishing pixel-level segmentation labels for data annotation. Figure 4 shows representative raw tongue images and their corresponding annotation results. The dataset is randomly divided into training, validation, and testing sets in a 6:2:2 ratio.

Table 1. Information of the tongue image datasets.

Dataset	Numbers	Marking Status	Data Sources
Dataset 1	300	With a marked	[16]
Dataset 2	1105	Not marked	Collected by professional (Our team)

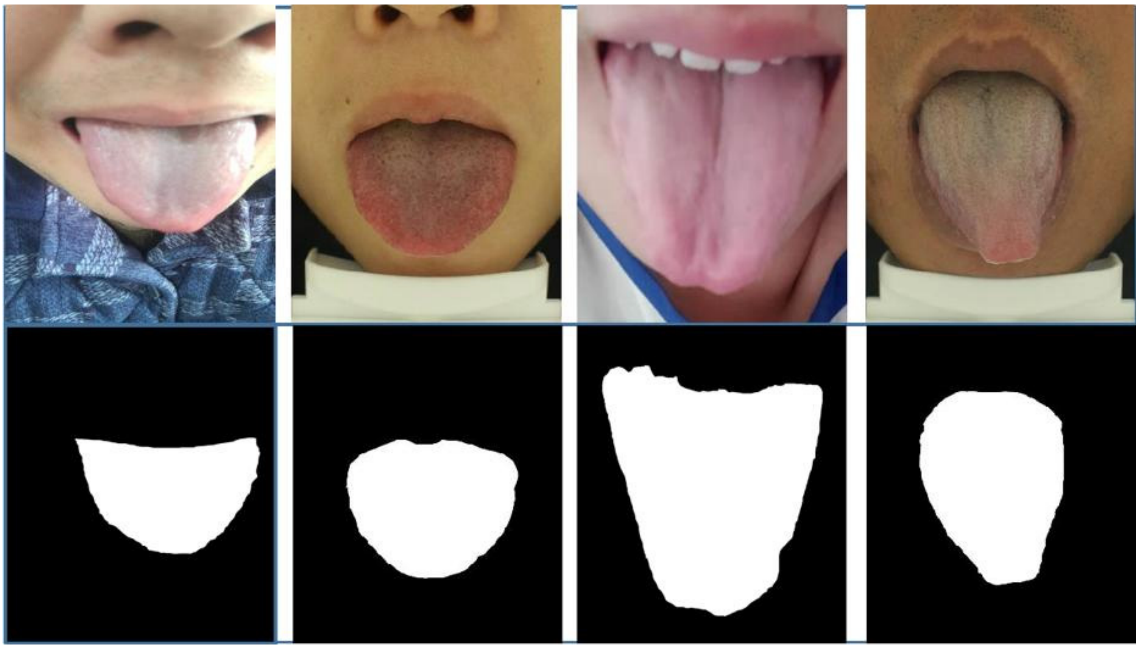

Figure 4. Some examples of tongue image segmentation dataset.

This study constructs a multi-stage data preprocessing system for the tongue image segmentation task. First, image size statistics revealed significant dimensional heterogeneity in the dataset. The bilinear interpolation algorithm standardizes all samples to a resolution of 416 × 352. Next, spatial domain filtering (Gaussian filter, bilateral filter) was applied to eliminate noise interference, and morphological opening operations were used to remove minor artifacts. Finally, based on a dynamic augmentation strategy, random rotations of ±15°, vertical flips, and HSV space parameter perturbations were introduced during the training phase to enhance the model's robustness to lighting variations.

3.2. Performance Evaluation Metrics

This study uses mean Intersection over Union (mIoU) and mean pixel accuracy (MPA) to evaluate the performance of different neural segmentation methods. mIoU is the ratio of the intersection and union points between the model's predicted results for each category and the true label values, and then the average is taken. Its definition is as follows.

$$mIoU = \frac{1}{K}\sum_{i=1}^{K}\frac{TP_i}{TP_i + FP_i + FN_i} \tag{1}$$

$$MPA = \frac{1}{K}\sum_{i=1}^{K}\frac{TP_i}{TP_i + FN_i} \tag{2}$$

where True Positive (TP) is the number of correctly segmented target pixels, True Negative (TN) is the number of correctly segmented background pixels, False Positive (FP) is the number of background pixels incorrectly segmented as target, and False Negative (FN) is the number of target pixels incorrectly segmented as background.

3.3. Experimental Environment and Configuration

The training and testing tasks for the tongue image segmentation model are implemented based on the PyTorch deep learning framework. The Adam optimizer is used to optimize the training process, and cross-entropy is employed as the loss function. The initial learning rate is set to 1 × 10^{-4}, halved every 10 iterations to promote model convergence and improve segmentation accuracy. The batch size is set to 16. These settings remain consistent across all models. Each model is trained under the same conditions, and the best-performing validation set weights are saved for subsequent testing.

3.4. Experimental Environment and Configuration

3.4.1. Comparisons with Other Methods

To evaluate the effectiveness of TSNet, we conducted a comparative analysis of its segmentation performance. The selected comparative models include FCN [17], SegNet [15], GCN [18], FusionNet [19], DeepLabv3+ [20], and BiSeNet [21].

Table 2 shows a performance comparison of different segmentation models, including mIoU (mean Intersection over Union), MPA (mean pixel accuracy), and the number of parameters. As shown in Table 2, the TSNet model excels in all evaluation metrics, particularly demonstrating its unique advantage in balancing accuracy and computational efficiency. Specifically, TSNet achieved the best mIoU of 97.20. The MPA is 98.52, only slightly lower than FusionNet. However, in practical applications, TSNet showcases excellent computational efficiency with a more reasonable parameter count (48.74 M). Compared to other models, such as FusionNet (81.68 M) and GCN (58.14 M), TSNet has fewer parameters, which allows it to maintain high accuracy while offering more substantial computational efficiency. In addition, TSNet performs better in balancing computational efficiency and performance compared to models like BiSeNet (mIoU: 93.54, MPA: 97.83, Parameters: 12.41 M) and DeepLabv3+ (mIoU: 96.05, MPA: 98.08, Parameters: 54.94 M). Particularly, TSNet demonstrates relatively excellent characteristics in balancing model size and accuracy. Its design allows it to provide efficient inference speeds across different hardware environments, while ensuring desirable segmentation accuracy.

Table 2. Performance comparison of different segmentation models. The bold values indicate the best results, and the underlined values indicate the second-best results.

Model	mIoU	MPA	Parameter
FCN	<u>96.80</u>	98.48	<u>20.22 M</u>
SegNet	96.54	98.30	29.46 M
GCN	94.71	97.60	58.14 M
FusionNet	96.72	**98.55**	81.68 M
DeepLabv3+	96.05	98.08	54.94 M
BiSeNet	93.54	97.83	**12.41 M**
TSNet	**97.20**	<u>98.52</u>	48.74 M

Figure 5 demonstrates the visual segmentation results of different methods, where TSNet shows a clear advantage, surpassing other existing tongue segmentation methods. In contrast, although other models can capture the shape of the tongue to some extent, they commonly suffer from issues such as blurred boundaries, over-segmentation, or loss of details. TSNet effectively avoids these problems, providing more accurate and consistent segmentation results, especially in handling the tongue's edges, where it excels in clearly defining the tongue's contours and accurately capturing details. In contrast, other models fall short in boundary definition and detail precision, often leading to over-segmentation or under-segmentation, particularly with complex or irregular tongue shapes. In summary, TSNet not only improves accuracy in the tongue segmentation task but also enhances robustness and boundary detail capture, demonstrating stronger adaptability.

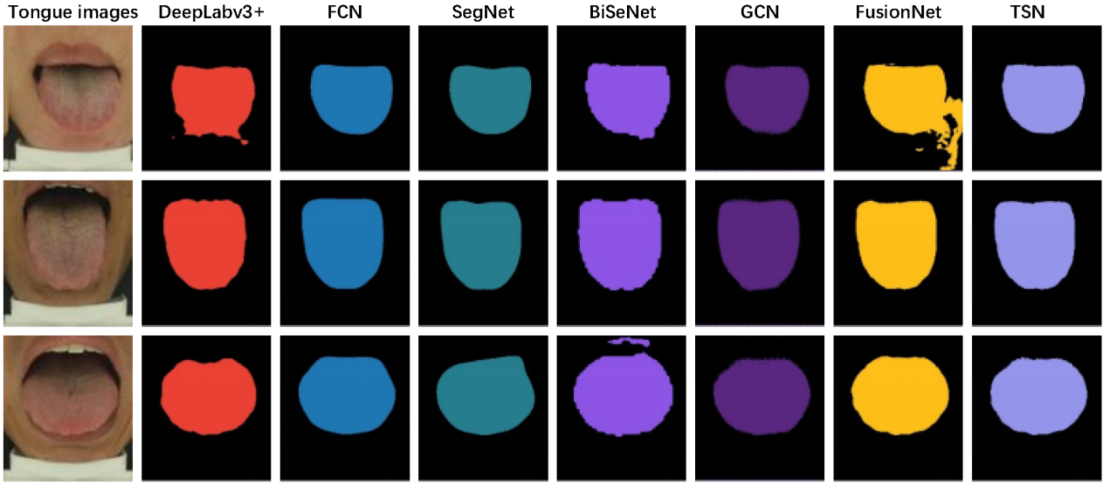

Figure 5. Visualization results of segmentation across different models.

3.4.2. Ablation Study

In this section, we conduct ablation studies to analyze the effectiveness of the proposed framework. Table 3 presents the different configurations of the model in the ablation experiments and their performance results.

Table 3. Ablation study results.

BA	SASPP	PASPP	mIoU	Parameter
-	-	-	95.89	8.48 M
√	-	-	96.58	39.82 M
√	√	-	96.74	47.16 M
√	-	√	97.20	48.74 M

According to the data in Table 3, when no additional modules (BA, SASPP, PASPP) are used, the model's mIoU is 95.89 with 8.48M parameters. After introducing the BA module, the mIoU increases to 96.58, but the parameter count significantly rises to 39.82 M, indicating that the BA module contributes to performance improvement. Further adding the SASPP module leads to a slight increase in mIoU to 96.74, but the parameter count increases again to 47.16 M. Finally, when the BA and PASPP modules are combined, the model's mIoU reaches the highest value of 97.20, with the parameter count at 48.74 M. This indicates that the combination of the BA and PASPP modules significantly enhances model performance, and the increase in parameters aligns with the performance improvement. As our model, the final configuration demonstrates an effective balance between high performance and computational cost, showcasing its effectiveness in the tongue segmentation task.

4. Discussions and Conclusions

This study proposes a new tongue image segmentation model, TSNet. The model integrates depthwise separable convolution, parallel atrous spatial pyramid pooling (PASPP), and boundary adjustment (BA) mechanisms, achieving a mIoU of 97.20% and an MPA of 98.52%, demonstrating outstanding segmentation performance. A key contribution of TSNet is its ability to strike a good balance between high segmentation accuracy and computational efficiency. Compared to other advanced models, such as FusionNet and GCN, TSNet achieves a higher mIoU score while maintaining a lower parameter count (48.74 M). The design of TSNet enables real-time performance even in resource-constrained environments and maintains high accuracy with minimal computational overhead.

In addition, the PASPP and BA modules proposed in this study not only demonstrate significant performance improvements within TSNet, but we also believe they hold strong potential for integration into other advanced backbone networks. Both modules feature modular and architecture-agnostic designs, making them suitable for embedding into mainstream segmentation networks such as UNet, DeepLabv3+, and SwinUNet. They are particularly effective in addressing common challenges in medical image segmentation, such as blurred boundaries and the need for multi-scale contextual awareness. Although we have not yet conducted a systematic evaluation of these modules in other architectures due to time and resource constraints, we plan to explore this direction in future work to assess their robustness and applicability across different tasks and networks, and to provide valuable insights for modular design in medical image segmentation.

Although TSNet has achieved promising results, some limitations can still be addressed in future work. First, the dataset used in this study contains 1405 tongue images, but it still does not fully represent the various shapes, colors, and conditions of tongue images. Future research could expand the dataset, particularly considering the tongue features across different populations, to improve the model's generalization capability. Secondly, TSNet faces issues with boundary recognition, mainly when the contrast between the tongue and the background is low or the edges are blurry. Future work could introduce boundary optimization mechanisms, such as boundary regression modules, conditional random field post-processing, and boundary loss functions, to improve segmentation accuracy and enhance the model's robustness in practical applications. In addition, although the PASPP and BA modules have shown strong potential within TSNet, their application in other backbone networks has not been fully explored. Future research will investigate the feasibility of integrating these modules into other mainstream networks to assess their generalizability and scalability.

In summary, TSNet provides a high-precision, computationally efficient solution for tongue image segmentation in Traditional Chinese Medicine, balancing segmentation accuracy and computational efficiency. By integrating methods such as depthwise separable convolution, PASPP, and BA, TSNet performs exceptionally well across various evaluation metrics, demonstrating its potential in the intelligentization of Traditional Chinese Medicine tongue diagnosis. Future research will focus on expanding the dataset, improving model efficiency, and exploring the application of TSNet in real-time diagnostic tools.

Author Contributions

All authors contributed to the study's conception and design. Material preparation and data collection were performed by D.X. and D.T. The first draft of the manuscript was written by D.X. and D.T. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Tian, D.; Chen, W.; Xu, D.; et al. A review of traditional Chinese medicine diagnosis using machine learning: Inspection, auscultation-olfaction, inquiry, and palpation. *Comput. Biol. Med.* **2024**, *170*, 108074.
2. Balasubramaniyan, S.; Jeyakumar, V.; Nachimuthu, D.S. Panoramic tongue imaging and deep convolutional machine learning model for diabetes diagnosis in humans. *Sci. Rep.* **2022**, *12*, 186.
3. Dong SU, I.; Zhang, L.; Fei, Y. Data-driven based four examinations in TCM: A survey. *Digit. Chin. Med.* **2022**, *5*, 377–385.
4. Wu, K.; Zhang, D. Robust tongue segmentation by fusing region-based and edge-based approaches. *Expert Syst. Appl.* **2015**, *42*, 8027–8038.
5. Cao, H.; Wang, Y.; Chen, J.; et al. Swin-unet: Unet-like pure transformer for medical image segmentation. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer Nature: Cham, Switzerland, 2022; pp. 205–218.
6. Zhu, Z.; Liu, L.; Free, R.C.; et al. OPT-CO: Optimizing pre-trained transformer models for efficient COVID-19 classification with stochastic configuration networks. *Inf. Sci.* **2024**, *680*, 121141.
7. Lu, S.Y.; Zhu, Z.; Tang, Y.; et al. CTBViT: A novel ViT for tuberculosis classification with efficient block and randomized classifier. *Biomed. Signal Process. Control* **2025**, *100*, 106981.
8. Lu, S.Y.; Zhang, Y.D.; Yao, Y.D. A regularized transformer with adaptive token fusion for Alzheimer's disease diagnosis in brain magnetic resonance images. *Eng. Appl. Artif. Intell.* **2025**, *155*, 111058.
9. Zhu, Z.; Ren, Z.; Lu, S.; et al. DLBCNet: A deep learning network for classifying blood cells. *Big Data Cogn. Comput.* **2023**, *7*, 75.

10. Huang, X.; Zhang, H.; Zhuo, L.; et al. TISNet-Enhanced fully convolutional network with encoder-decoder structure for tongue image segmentation in Traditional Chinese Medicine. *Comput. Math. Methods Med.* **2020**, *2020*, 6029258.
11. Yao, L.; Xu, Y.; Zhang, S.; et al. HPA-UNet: A Hybrid Post-Processing Attention U-Net for Tongue Segmentation. *IEEE J. Biomed. Health Inform.* **2024**. https://doi.org/10.1109/JBHI.2024.3446623
12. Jia, G.; Cui, Z.; Fei, Q. QA-TSN: QuickAccurate Tongue Segmentation Net. *Knowl. -Based Syst.* **2025**, *307*, 112648.
13. Huang, Z.; Huang, R.; Zhang, J.; et al. Attention guided tongue segmentation with geometric knowledge in complex environments. *Biomed. Signal Process. Control* **2025**, *104*, 107426.
14. Chen, L.C.; Papandreou, G.; Kokkinos, I.; et al. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848.
15. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495.
16. BioHit. Tongueimagedataset. 2014. Available online: https://github.com/BioHit/TongeImageDataset (accessed on 1 August 2024).
17. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
18. Peng, C.; Zhang, X.; Yu, G.; et al. Large kernel matters—Improve semantic segmentation by global convolutional network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4353–4361.
19. Quan, T.M.; Hildebrand, D.G.C.; Jeong, W.K. Fusionnet: A deep fully residual convolutional neural network for image segmentation in connectomics. *Front. Comput. Sci.* **2021**, *3*, 613981.
20. Chen, L.C.; Zhu, Y.; Papandreou, G.; et al. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
21. Yu, C.; Wang, J.; Peng, C.; et al. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 325–341.