IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received September 14, 2019, accepted September 26, 2019, date of publication October 11, 2019,
date of current version October 24, 2019.

*Digital Object Identifier 10.1109/ACCESS.2019.2946681*

# Tonguenet: Accurate Localization and Segmentation for Tongue Images Using Deep Neural Networks

**CHANGEN ZHOU**[1], **HAOYI FAN**[2,3], **AND ZUOYONG LI**[1,3]

[1]College of Traditional Chinese Medicine, Fujian University of Traditional Chinese Medicine, Fuzhou 350122, China
[2]School of Computer Science and Technology, Harbin University of Science and Technology, Harbin 150080, China
[3]Fujian Provincial Key Laboratory of Information Processing and Intelligent Control, Minjiang University, Fuzhou 350121, China

Corresponding author: Zuoyong Li (fzulzytdq@126.com)

This work was supported in part by the National Natural Science Foundation of China under Grant U1705286, Grant 61866005, and Grant 61972187, in part by the Research and Development of Internet-based Traditional Chinese Medicine Case Collection Platform under Grant JG2017016, in part by the Fujian Province 2011 Collaborative Innovation Center of TCM Health Management, in part by the Key Project of College Youth Natural Science Foundation of Fujian Province under Grant JZ160467, and in part by the Fujian Provincial Leading Project under Grant 2017H0030.

**ABSTRACT** Tongue diagnosis is an important way of monitoring human health status in traditional Chinese medicine. As a key step of achieving automatic tongue diagnosis, the major challenges for robust and accurate segmentation and identification of tongue body in tongue images lay in the large variations of tongue appearance, e.g., tongue texture and tongue coating, caused by different diseases for different patients. To cope with these challenges, we propose a novel end-to-end model for multi-task learning of tongue localization and segmentation, named TongueNet, in which pixel-level prior information is utilized for supervised training of deep convolutional neural network. Firstly, we introduce a feature pyramid network based on the designed context-aware residual blocks for the extraction of multi-scale tongue features. Then, the region of interests (ROIs) of tongue candidates are located in advance from the extracted feature maps. Finally, finer localization and segmentation of tongue body are conducted based on the feature maps of ROIs. Quantitative and qualitative comparisons on real-world datasets show that the proposed TongueNet achieves state-of-the-art performance for the segmentation of tongue body in terms of both robustness and accuracy.

**INDEX TERMS** Tongue segmentation, tongue diagnosis, traditional Chinese medicine, TongueNet, deep learning.

## I. INTRODUCTION

Tongue diagnosis is one of the most common and important methods used in Traditional Chinese Medicine (TCM) because of its painless attribute and convenience. The appearance features of tongue body such as color, texture, shape, and coating reveal a large amount of information about human health status in TCM. However, traditional tongue diagnosis depends highly on clinicians' experience and thus different clinicians are likely to conclude different diagnostic results for the same patient. Fortunately, by utilizing computer and other relative techniques, computer-aided methods for tongue diagnosis are able to improve these deficiencies [1], [2]. In those methods, the tongue ROI extraction process is firstly required to analyze the tongue appearance features automatically. Therefore, the automatic localization and segmentation of tongue body in tongue images is one of the key technologies for tongue diagnosis and the performance of which directly affects the accuracy of tongue diagnosis.

Tongue segmentation aims at segmenting the region of tongue body from those complex backgrounds such as teeth, lips, face and other materials. Different from the conventional segmentation tasks in nature scene, tongue segmentation is more challenging because of the issues of large variations of tongue appearances for different patients while higher precision requirement, data imbalance, e.g. small parts of foreground region (tongue body) compared with the background region, and hard sample mining e.g. lip pixels as the hard samples is hard to be segmented from tongue pixels because the similar appearances and close touch between them.

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

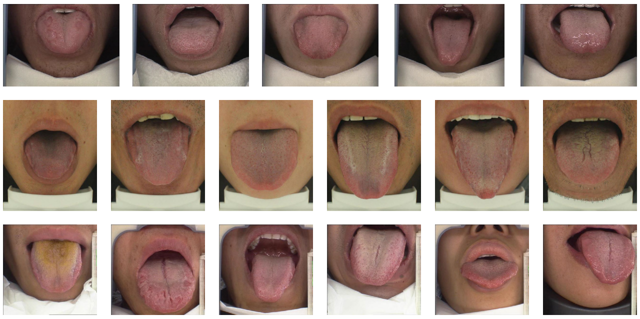


**FIGURE 1.** Examples of tongue images with large variations of tongue appearances from different patients.

Different methods have been proposed in the past for the segmentation of tongue. Most of those methods are based on the traditional image processing techniques, however, some of them are sensitive to the illumination changes or clustered backgrounds [3], [4], some of them confuse lips from tongue body [5]–[7], and some of them require additional preprocessing which makes the whole segmentation process more complex [8], [9]. More recently, deep learning based methods [10]–[12] have been proposed for automatic tongue segmentation. Although those deep learning based methods outperform the most traditional tongue segmentation methods, there are still some limits in those methods. In [10], additional preprocessing is required such as image enhancement which makes the whole segmentation process more complex. Similarly, in [11], brightness discrimination as a preprocessing reduces the ability of generalization as a deep learning based model. In [12], an end-to-end method is proposed and achieves a remarkable result in tongue segmentation based on SharpMask [13], however, SharpMask is capable of locating the object but can not distinguish the type of object, which results in slow segmentation speed and less accurate because of undesired processing on unrelated objects.

To solve above-mentioned issues, we propose a novel end-to-end model for multi-task learning of tongue localization and segmentation, called TongueNet, which segments tongue in a pixel-to-pixel manner automatically based on deep convolution neural network and achieves high accuracy. To extract discriminative multi-scale features of tongue images, a novel feature pyramid network based on the designed context-aware residual blocks is proposed. Then, based on the extracted multi-scale feature maps, TongueNet is able to localize the tongue body rapidly, and then generates a precise segmentation mask of tongue body, which does not require any preprocessing and is robust to shape deformation, color variations of tongue coating and different illumination conditions. Moreover, a weighted loss function based on Tversky index is utilized for the training of the model, to mitigate the issues of data imbalance and hard sample mining.

The main contributions of this paper are as follows:

- We propose an end-to-end deep model, named TongueNet, which segments tongue in a pixel-to-pixel manner. Different from previous deep learning based methods [10]–[12], which segment the tongue body directly on the original image with a complex background. In TongueNet, the candidate regions of tongue ROI are located firstly before segmentation, which greatly reduces the difficulty of segmentation and therefore boosts the segmentation accuracy.
- We develop a context-aware feature pyramid network based on dilated residual convolution [14], which is designed according to the characters of tongue image, to extract multi-scale features of tongue effectively.
- We conduct quantitative and qualitative comparisons among TongueNet and the state-of-the-art methods on the commonly used datasets, and the segmentation results indicate that TongueNet outperforms other methods significantly.

## II. RELATED WORK

### A. TRADITIONAL METHODS

Traditional tongue segmentation methods based on images features can be roughly divided into three categories: region based methods [3], [4], [9], edge based methods [5]–[7], [15], and the hybrid methods [8]. For region based methods, in [3], a tongue segmentation method based on the combination of the watershed transform and active contour model is proposed, in which, the watershed transform is used to get the initial contour, and an active contour model, or "snakes", is used to converge to the precise edge. Similarly, in [4], the watershed algorithm is utilized to segment tongue image into many small regions and then color-similarity based region merging is performed to get the final tongue body segmentation. In [16] histogram projection is utilized firstly to locate the tongue body, getting the convinced foreground and background area in form of trimap. Then, trimap is took as the input for LBDM [17] algorithm to implement the final segmentation. In [9], the initial region of tongue body is firstly determined via transforming and thresholding on the hue component of HSI color model and the red component of RGB color model, and then refined by removing fake object regions such as the upper lip with the help of morphological operations. For edge based methods, edge initialization are required in those methods to segment the final tongue body. In [5], the evolving contour is initialized with the bi-elliptical deformable template. In [15], the initial edge of tongue body is selected from edge map. Moreover, there are also some color feature based methods for initial contour localization [6], [7]. As a hybrid method that fuses region-based and edge-based approaches into one single segmentation pipeline [8], the ROI of tongue body is firstly extracted based on the use of color information, and then the original image is replaced by the ROI for subsequent segmentation. Finally an improvement is made to an existing region-based method MSRM (Maximal Similarity based Region Merging). Although each of those methods has its fair share of success, corresponding limitations still exist

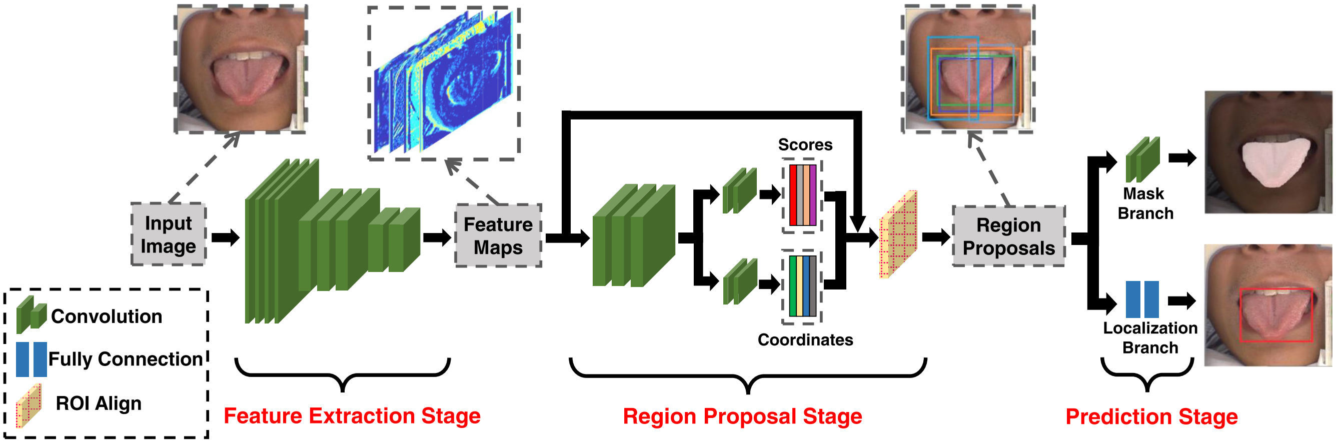


**FIGURE 2.** The framework of the proposed TongueNet.

among them. For example, some of them are sensitive to the illumination changes and the cluttered backgrounds, some of them confuse the lips from tongue body and therefore some preprocessings are required, which is time-consuming. So, it is still considerably challenging to achieve sufficient efficiency and accuracy for tongue segmentation.

### B. DEEP LEARNING BASED METHODS

Benefiting from the development of deep learning based techniques, such as deep convolution neural networks, significant performance improvements have been witnessed in the field of computer vision; e.g. object detection [18], image classification [19], and image segmentation [20], [21]. However, few works employ deep learning-based methods for tongue segmentation due to the difficulty of collecting and labeling the tongue images. Recently, several deep learning based methods [10]–[12] have been proposed for automatic tongue segmentation, which outperform most of the traditional tongue segmentation methods. In [10], an enhanced HSV color model convolutional neural network is proposed for tongue image segmentation, however, in this model, additional preprocessing is required such as image enhancement which makes the whole segmentation process more complex. In [11], an image quality evaluation method based on brightness statistics is proposed to judge whether the input image is to be segmented, and then SegNet [22] is employed for training on the preprocessed images, in this method, brightness discrimination as a preprocessing reduces the ability of generalization as a deep learning based model. In [23], initial segmentation results of Deeplabv3 [24] are firstly obtained and then optimized by LBDM [17] to get the final segmentation results. In [12], an end-to-end deep model is proposed based on ResNet [25] and SharpMask [13], however, in this method, segmentation precedes recognition, which is slow and less accurate.

### C. MULTI-TASK LEARNING

Multi-task learning [26] aims at learning multiple tasks jointly by exploiting the shared structures to improve generalization performance and mitigate manual labeling consumption [27]. Recently, multi-task learning has achieve remarkable success in the field of computer vision such as image classification [28], semantic segmentation [29], image-to-image prediction [30], and depth prediction [31]. In Uber-Net [32], an image pyramid approach is proposed to process images across multiple resolutions, where for each resolution, additional task-specific layers are constructed on the top of the shared VGG-Net [33]. Mask R-CNN proposed in [29] detects objects in an image while simultaneously generating a high-quality segmentation mask for each instance. Mask R-CNN extends Faster R-CNN by adding a branch for predicting an object mask in parallel with the existing branch for bounding box recognition. In [31], Cross-Stitch Networks contain one standard feed-forward network per task, with cross-stitch units to allow features to be shared across tasks. Inspired by the successful application of multi-task learning architecture in computer vision, our proposed TongueNet is a multi-task learning based model that localizes and segments tongue body simultaneously.

## III. METHOD

For the segmentation of tongue body in tongue images, we aim to automatically segment the tongue body from the complex background, independent of the diversity in their appearance, and without any manual intervention and preprocessing. To achieve this, we model the problem as a binary dense labeling task: Given a camera-taken RGB image, which contains both tongue body and other noise background regions such as face and lip, the task is to predict either "tongue" or "non-tongue" labels for each pixel.

As a multi-task learning architecture, our proposed TongueNet consists of three stages, namely feature extraction stage, region proposal stage, and prediction stage, to conduct tongue localization and segmentation tasks simultaneously. Firstly, to make full use of spatial information and prior knowledge such as color, shape, and texture of tongue body, a context-aware feature pyramid network based on dilated

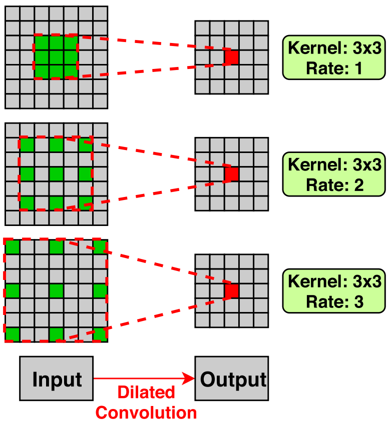


**FIGURE 3. Examples of 3 × 3 dilated convolutional kernels with three different dilation rates as 1, 2, and 3 respectively.**

residual convolution is employed in the feature extraction stage to extract multi-scale features of tongue images. Then, tongue ROIs are localized based on the extracted feature maps in the region proposal stage, which are served as the candidates for the final localization and segmentation of tongue body. Finally, in the prediction stage, two different learning tasks, namely segmentation and localization respectively, are jointly learned by TongueNet via optimizing the designed loss function. The details of the proposed method are described in the following sections.

## A. CONTEXT-AWARE FEATURE PYRAMID NETWORKS

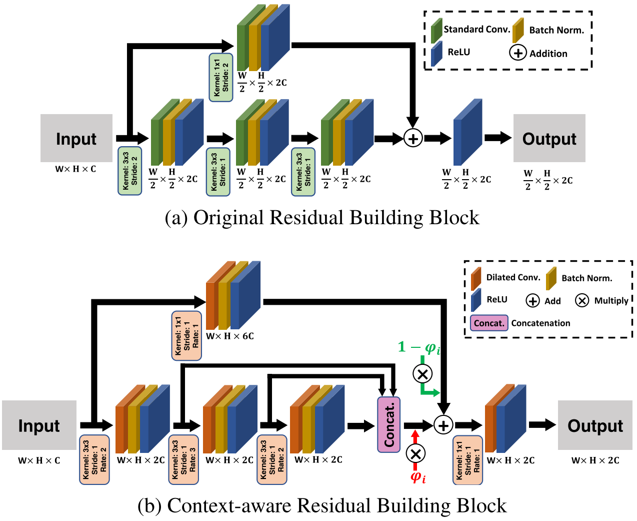
### 1) DILATED CONVOLUTION

Dilated convolution [34], [35] aims at enlarging the receptive field of feature maps to aggregate context information without increasing extra parameters and computation, e.g., adding more convolutional layers enables a larger receptive field but introduce more filtering operations and thus more computation. Different from standard convolution operation, in which, the striding operation and pooling layers would lead to a reduction of the resolution of feature maps and therefore result in a loss of spatial information, in dilated convolution operation, there are not above-mentioned issues.

Mathematically, a 2-D dilated convolutional filtering can be formulated as follows:

$$y(h, w) = \sum_{i=1}^{H} \sum_{j=1}^{W} x(h + r \times i, w + r \times j) f(i, j) \tag{1}$$

where $y(h, w)$ is the output of dilated convolutional filtering from input $x(h, w)$ at location $(h, w)$, $f(i, j)$ is a kernel with the height of $H$ and the width of $W$ respectively, and $r$ is the dilation rate. Therefore, a size of $k \times k$ kernel is capable of filtering a size of $(k + (k - 1)(r - 1)) \times (k + (k - 1)(r - 1))$ region with dilated rate $r$. For example, as shown in Fig. 3, a size of $3 \times 3$ dilated convolutional kernel with rate 1 is the standard convolutional kernel, whose size of receptive field is $3 \times 3$, while the sizes of receptive field of a size of $3 \times 3$ dilated convolutional kernel are $5 \times 5$ for rate 2 and $7 \times 7$ for rate 3 respectively.



(a) Original Residual Building Block

(b) Context-aware Residual Building Block

**FIGURE 4. Comparison of the original residual building block and the proposed context-aware residual building block.**

### 2) CONTEXT-AWARE DILATED RESIDUAL BLOCK

As illustrated in many previous works [33], [36], a good feature extraction network should be deep enough with many convolution layers such that multi-scale features can be sufficiently learned. Inspired by the successful use of ResNet [25] in feature extraction and classification tasks, residual blocks are also employed in our proposed TongueNet but with a context-aware modification to aggregate more context information of tongue body for more discriminative feature extraction.

The original residual block contains several convolutional groups with different kernel sizes and the mapping performed by the original residual block can be defined as follows:

$$x_{i+1} = ReLU(G(x_i, w_i^g) + F(x_i, w_i^f)) \tag{2}$$

where $x_i$ and $x_{i+1}$ are the input and output of $i$-th residual block respectively, $G(\cdot)$ and $F(\cdot)$ are two different nonlinear transformation groups, each of which consists of a convolution, batch normalization (BN) and rectified linear units (ReLU). Especially, $F(\cdot)$ is usually one single nonlinear transformation group or an identity function $I(x_i) = x_i$. $w_i^g$ and $w_i^f$ are two sets of weights and biases associated with $G(\cdot)$ and $F(\cdot)$ respectively. For example, as shown in Fig. 4a, given a 3-groups based original residual block, with each group has a batch normalization layer and a ReLU layer followed by a standard convolutional layer, the resolution of output is half of input, which might lead to a loss spatial information for feature extraction.

Different from the original residual block, our proposed context-aware dilated residual block is defined as follows:

$$x_{i+1} = ReLU(D(\varphi_i \times G_D(x_i, w_i^g) + (1 - \varphi_i) \times F_D(x_i, w_i^f))) \tag{3}$$

where $x_i$ and $x_{i+1}$ are the input and output of $i$-th context-aware residual block respectively, $D$ is the dilated convolution

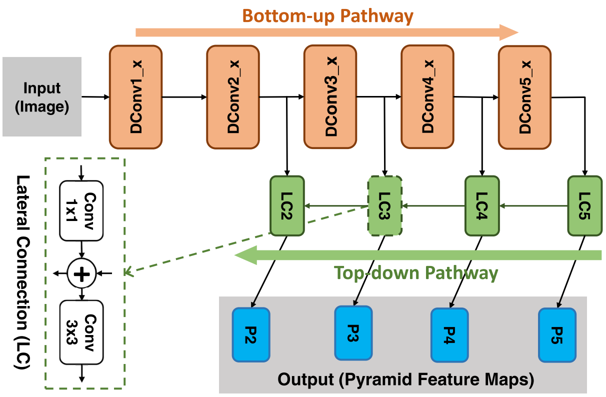


**FIGURE 5.** The architecture of the proposed context-aware feature pyramid networks.

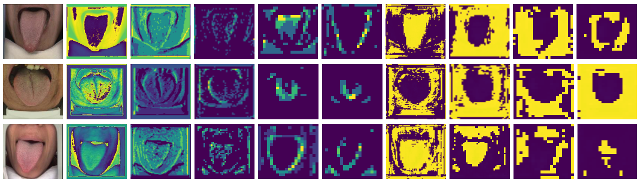


**FIGURE 6.** Examples of feature maps extracted from context-aware feature pyramid networks. Column 1-10 are the original RGB image, outputs of DConv1_x, DConv2_x, DConv3_x, DConv4_x, DConv5_x, P2, P3, P4, and P5 respectively; Row 1-3 are three different tongue samples from three different datasets. (The feature maps at different levels are randomly selected for visualization.).

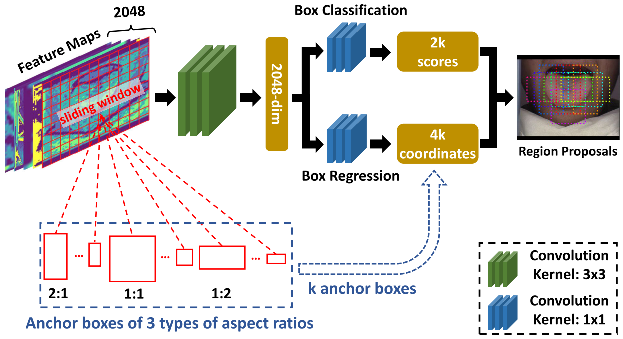


**FIGURE 7.** Region proposal network used in TongueNet.

operation, $G_D(\cdot)$ and $F_D(\cdot)$ are two different nonlinear transformation groups, each of which consists of a dilated convolution, batch normalization (BN) and rectified linear units (ReLU). $w_i^g$ and $w_i^f$ are two sets of weights and biases associated with $G_D(\cdot)$ and $F_D(\cdot)$ respectively. Here, a weighted skip connection is employed by assigning different weights $\varphi_i$ and $1-\varphi_i$ to the outputs of $G_D(\cdot)$ and $F_D(\cdot)$ respectively.

### 3) MULTI-SCALE FEATURE EXTRACTION

Although the powerful representation ability that deep residual network possesses, it has been proven that a better performance can be boosted through using pyramid representations for multi-scale images [37]. Therefore, a feature pyramid network (FPN) based on the proposed context-aware residual blocks is employed in TongueNet to extract more reliable and representative multi-scale features. Fig. 5 shows the detailed architecture of the proposed context-aware FPN, which consists of three parts: the bottom-up pathway, the lateral connections, and the top-down pathway.

In the bottom-up pathway, five context-aware dilated residual blocks, namely DConv1_x, DConv2_x, DConv3_x, DConv4_x, and DConv5_x respectively, are serially connected as the backbone of FPN for feature extraction of tongue region. The output of each block is employed to construct the corresponding feature maps in different pyramid levels. As we go down the top-down pathway through a module called lateral connection, feature maps are firstly merged with the corresponding bottom-up features by going through a 1 × 1 convolution layer with an element-wise addition operator. Then, we apply a 3 × 3 convolution again to all merged feature maps to obtain the final pyramid feature maps, denoted as P2, P3, P4, P5. The 3 × 3 convolution used here is to reduce the aliasing effect due to upsampling. Each level of the pyramid feature maps can be used for tongue body localization at a different scale in the next region proposal stage as shown in Fig. 2.

## B. ARCHITECTURE OF TONGUENET

Our proposed TongueNet consists of three stages for tongue localization and segmentation, namely feature extraction stage, region proposal stage, and prediction stage. The architecture of TongueNet is illustrated in Fig. 2.

In the feature extraction stage, a context-aware feature pyramid network is utilized to extract multi-scale features of tongue images, as described in section III-A.3. In this stage, the feature maps of tongue image in different pyramid levels are extracted for tongue localization in the next region proposal stage. As shown in Fig. 6, the extracted feature maps in different levels represent discriminatively for tongue body compared to the face and other substances.

In the region proposal stage, tongue ROIs are localized for the final localization and segmentation of tongue body. In this stage, the region proposal network (RPN) [38] is employed, which slides across the multi-scale feature maps generated from the lower feature extraction module by a sliding window, to produce a list of region proposals that likely contain tongue body. Given those feature maps as input, the output of RPN is a series of region proposals that might contain the tongue goal. As shown in Fig. 7, the feature maps are passed to a 3 × 3 size of standard convolutional layer and then mapped to a 2048-dimensional vector. Follow that, two branches of 1 × 1 size of standard convolutional layers, namely box classification branch and box regression branch respectively, are then used to predict the category and position of a proposal, which is also called an anchor and centered at the corresponding sliding window as depicted as a red box in Fig. 7 for visualization. For each sliding-window, both branches simultaneously predict $k$ proposals, in which three different anchor aspect ratios {2:1, 1:1, 1:2} and four scales determined by four levels of feature maps {P2, P3, P4, P5} are employed to generate $k = 12$ anchors in this case at each sliding position. Finally, $2k$ scores that measure the probability of tongue/non-tongue for each proposal, and $4k$ outputs that encode the coordinates of $k$ proposal boxes, are obtained from box classification branch and box regression branch respectively.

In the prediction stage, each of the proposals generated from the region proposal stage by sliding over the feature maps, is converted into the fixed size of feature map by using bilinear interpolation through a technique called RoIAlign [29], to rectify the misaligned tongue proposals. After alignment, two followed branches are applied, namely localization branch, and mask branch, to perform final tongue body localization and tongue mask segmentation respectively, as demonstrated in Fig. 2.

In the localization branch, two concatenated FC layers as a regressor to refine the anchor box to fit the tongue object better for proposal localization. In the mask branch, which consists of a few stacked convolutional layers, the feature maps selected by RPN are taken as inputs, and a pixel-to-pixel semantic segmentation is conducted based them to get the final segmentation mask of tongue body.

### C. LOSS FUNCTION

As in many multi-task learning models [28]–[31], a multi-task loss function is utilized in this paper for the training of TongueNet, as defined in the follows:

$$L = L_{loc} + L_{mask} \tag{4}$$

where $L_{loc}$ is the sum of a smooth L1 loss and a log loss [39] of the bounding box generated by the RPN in the region proposal stage of TongueNet. $L_{mask}$ is a weighted loss function based on Tversky index [40] for hard sample mining, e.g. lip pixels are the hard samples which are similar with tongue body pixels and touch to them closely, and data imbalance, e.g. the area of tongue body is far small compared with background area. The definition of $L_{mask}$ is as follows:

$$L_{mask} = \sum_{c} 1 - TI_c \tag{5}$$

where $TI_c$ is the Tversky similarity score [41] and defined as follows:

$$TI_c = \frac{\sum_{i=1}^{N} p_{ic} g_{ic} + \epsilon}{\sum_{i=1}^{N} p_{ic} g_{ic} + \alpha \sum_{i=1}^{N} p_{ic} g_{i\bar{c}} + \beta \sum_{i=1}^{N} p_{i\bar{c}} g_{ic} + \epsilon} \tag{6}$$

where $p_{ic}$ is the probability that pixel $i$ belong to tongue class $c$, $p_{i\bar{c}}$ is the probability that pixel $i$ belong to non-tongue class $\bar{c}$, $g_{ic}$ is the ground truth training label which is 1 for tongue pixel $i$ and $g_{i\bar{c}}$ is 0 for non-tongue pixel $i$. $\epsilon$ is a small number set as $10^{-8}$ by us to prevent division by zero. $\alpha$ and $\beta$ are two parameters which control the trade off between false negative and false positive in the case of large class imbalance. When $\alpha = \beta = 0.5$, $TI_c$ is simplified to a Dice score coefficient. In this paper, $\alpha$ is set as 0.3 and $\beta$ 0.7 empirically for all experiments.

## IV. EXPERIMENT

In this section, we discuss the experiments to validate the effectiveness of the proposed architecture. Firstly, we introduce the dataset and evaluation method. Then, experimental results are demonstrated which prove that TongueNet not only achieves a state-of-the-art segmentation accuracy on the commonly used datasets, but also performs stable on the newly collected complex dataset.

### A. DATASET AND EVALUATION METHOD

For medical tongue segmentation, the existed public datasets are small and the tongue images in these datasets are captured in a uniform illumination condition and small changes of tongue body in appearance. To evaluate the proposed method in term of both accuracy and robustness, we use three different tongue image datasets, called TestSet1, TestSet2, TestSet3 respectively, among which, TestSet1 [12] is a common dataset containing 300 images with size of 768 × 576 published by BioHit;[1] TestSet2 was collected by us from by the Third People's Hospital of Fujian Province which consists of 331 images with size of 550 × 650, the tongue images in this dataset were captured from hundreds of patients and most of them differ greatly in terms of shape deformation, color of tongue coating and tongue textures; and TestSet3 [8], [42] contains totally 290 images with size of 600 × 576, in which, the images are captured by the same device but under two different cases: whether to take the tongue lateral view or not. For TonguSet3, we crop the images and use only the front view parts for the experiment.

For evaluation metrics, three commonly used performance measures in deep learning based methods [11], [12], namely Precision, Dice coefficient (Dice) and mean Intersection over Union (mIoU) respectively, and other three commonly used metrics for traditional segmentation methods [5], namely False Positive Rate (FPR), False Negative Rate (FNR) and Misclassification Error (ME) respectively. All six metrics are defined as follows:

$$Precision = \frac{|F_g \bigcap F_p|}{|F_p|} \tag{7}$$

$$Dice = \frac{2|F_g \bigcap F_p|}{|F_g| + |F_p|} \tag{8}$$

$$mIoU = \frac{1}{2}\left(\frac{|F_g \bigcap F_p|}{|F_g \bigcup F_p|} + \frac{|B_g \bigcap B_p|}{|B_g \bigcup B_p|}\right) \tag{9}$$

$$FPR = \frac{|B_g \bigcap F_p|}{|B_g|} \tag{10}$$

$$FNR = \frac{|F_g \bigcap B_p|}{|F_g|} \tag{11}$$

$$ME = 1 - \frac{|F_g \bigcap F_p| + |B_g \bigcap B_p|}{|F_g| + |B_g|} \tag{12}$$

where $F_p$ and $B_p$ are tongue region (foreground) and non-tongue region (background) of the prediction of model, respectively; $F_g$ and $B_g$ are tongue region (foreground) and

[1] https://github.com/BioHit/TongeImageDataset

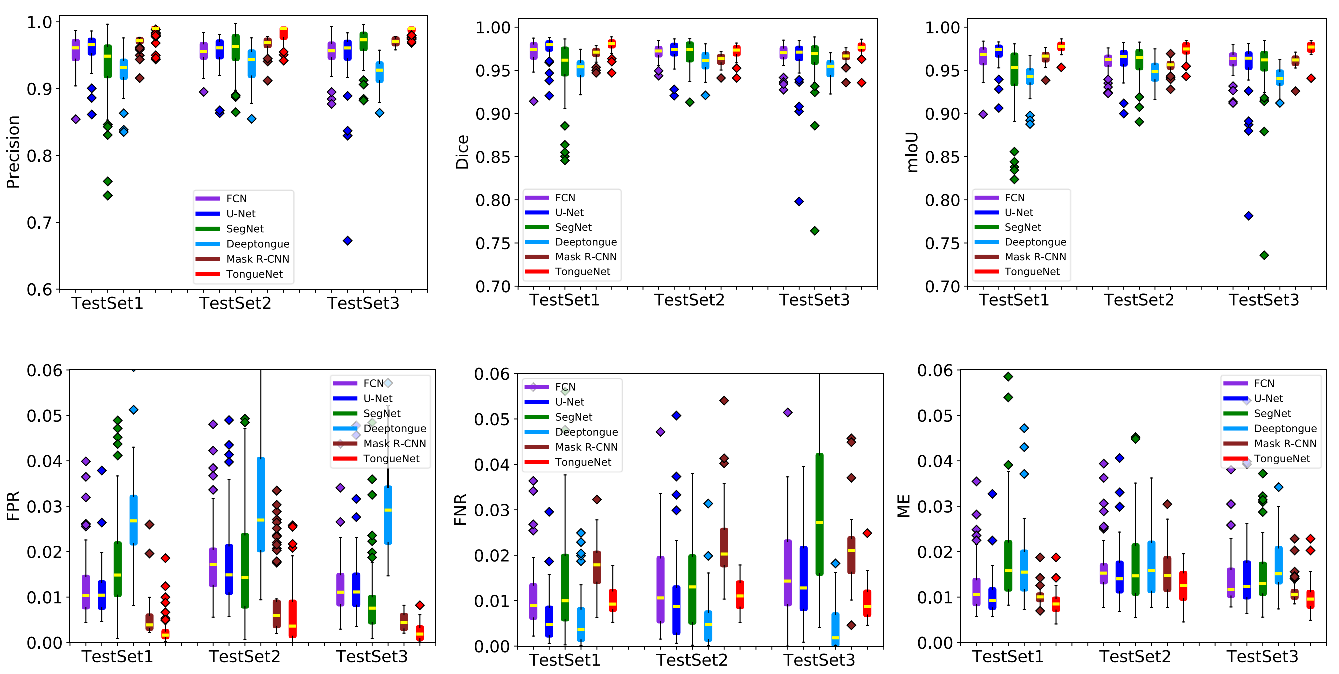


**FIGURE 8.** Box-and-whisker plots of different metrics for performance comparison among TongueNet and other baselines.

non-tongue region (background) of the ground truth, respectively; $|\cdot|$ is the cardinality of a set.

## B. IMPLEMENT DETAILS

The model is implemented by using Keras[2] deep learning framework and trained on Ubuntu 16.04 OS with 2.5GHz Intel Core i7 CPU, 32GB RAM, and NVIDIA GTX 1080Ti graphic card with 11GB memory. Adam algorithm [43] is utilized for optimization with learning rate as 0.001, batch size 4, weight decay $2 \times 10^{-4}$. The training set, validation set, and test set are produced by randomly splitting 80%, 10% and 10% of each dataset respectively for training and testing. In the experiment, ten-fold cross-validation is utilized for each model to evaluate the performance. During training, data augmentations such as scale, rotation, and changes on brightness, contrast, gamma, and color are performed as in other deep learning based methods [11], [12].

## C. EXPERIMENTAL RESULTS AND ANALYSIS

We compare our methods with five of the most recent deep learning based methods: FCN [44], U-Net [45], Mask R-CNN [29], SegNet based method proposed in [11], DeepTongue [12]. It should be noted that FCN, U-Net, and Mask R-CNN are three baselines that were originally used for segmentation in nature scene or single task medical image segmentation. Different from [11], [12], which selected only the mean value of variant metrics to evaluate the segmentation performance of tongue images, in this paper, box-and-whisker plots are also utilized to analyze the distribution of the results.

### 1) QUANTITATIVE RESULTS

To demonstrate the effectiveness of the proposed method quantitatively, six metrics precision, dice score, mIoU, FPR, FNR, and ME are selected for the performance measurement, and the results for all compared methods on Testset1, Testset2, Testset3 datasets are provided in Table 1. and Fig. 8.

The experimental results show that the proposed TongueNet significantly outperforms the baselines on all datasets. For example, as shown in Table 1, TongueNet achieves gains of 1.45% compared with FCN, 0.75% compared with U-Net, 3.58% compared with SegNet, 3.57% compared with DeepTongue, and 1.22% compared with Mask R-CNN respectively on TestSet1 for mIoU metric. On TestSet2, TongueNet decreases the ME value by around 0.41% compared with FCN, 0.31% compared with U-Net, 0.43% compared with SegNet, 0.51% compared with DeepTongue, and 0.11% compared with Mask R-CNN, respectively. All those improvements demonstrate the effectiveness of multi-task learning and context-aware based feature extraction module in TongueNet. However, for FNR, it seems that the performance of TongueNet is slightly worse than those compared baselines, for example, TongueNet increases the FNR value by around 0.51% compared with U-Net on TestSet1, 0.24% compared with DeepTongue on TestSet2, and 0.26% compared with DeepTongue on TestSet3, the true cause for this is because the predictions by those baselines are not closed enough to the real boundary of

[2] https://github.com/keras-team/keras

**TABLE 1. Segmentation performance of different methods. Best results are marked in red and the second-best in blue.**

| Dataset | Method | Precision | Dice | mIoU | FPR | FNR | ME |
|---|---|---|---|---|---|---|---|
| TestSet1 | FCN | 95.26% | 97.01% | 96.29 % | 1.35% | 1.14% | 1.29% |
| | U-Net | 95.86% | **97.58%** | **96.99%** | 1.12% | **0.6%** | 1.02% |
| | SegNet | 92.88% | 95.16% | 94.16% | 2.03% | 2.1% | 2.09% |
| | DeepTongue | 92.78% | 95.22% | 94.17% | 2.71% | 1.17% | 1.69% |
| | Mask R-CNN | **96.97%** | 96.98% | 96.52% | **0.51%** | 1.7% | **0.96%** |
| | TongueNet | **98.63%** | **97.96%** | **97.74%** | **0.26%** | **1.01%** | **0.85%** |
| TestSet2 | FCN | 95.47% | 97.02% | 95.94% | 1.85% | 1.35% | 1.66% |
| | U-Net | 95.54% | **97.23%** | **96.23%** | 1.79% | **0.98%** | 1.56% |
| | SegNet | 95.63% | 96.95% | 95.89% | 1.75% | 1.61% | 1.68% |
| | DeepTongue | 93.77% | 96% | 94.79% | 3.09% | **0.87%** | 1.76% |
| | Mask R-CNN | **96.67%** | 96.27% | 96.04% | **0.93%** | 2.17% | **1.45%** |
| | TongueNet | **98.28%** | **97.48%** | **97.18%** | **0.63%** | 1.1% | **1.25%** |
| TestSet3 | FCN | 95.45% | **96.84%** | 96.06% | 1.28% | 1.69% | **1.37%** |
| | U-Net | 94.83% | 96.32% | 95.51% | 1.46% | 1.95% | 1.58% |
| | SegNet | 96.61% | 96.12% | 95.3% | 0.97% | 4.1% | 1.65% |
| | DeepTongue | 92.39% | 95.16% | 94.11% | 2.92% | **0.7%** | 1.7% |
| | Mask R-CNN | **97.32%** | 96.65% | **96.17%** | **0.45%** | 2.1% | 1.48% |
| | TongueNet | **98.71%** | **97.61%** | **97.16%** | **0.24%** | **0.96%** | **1.01%** |

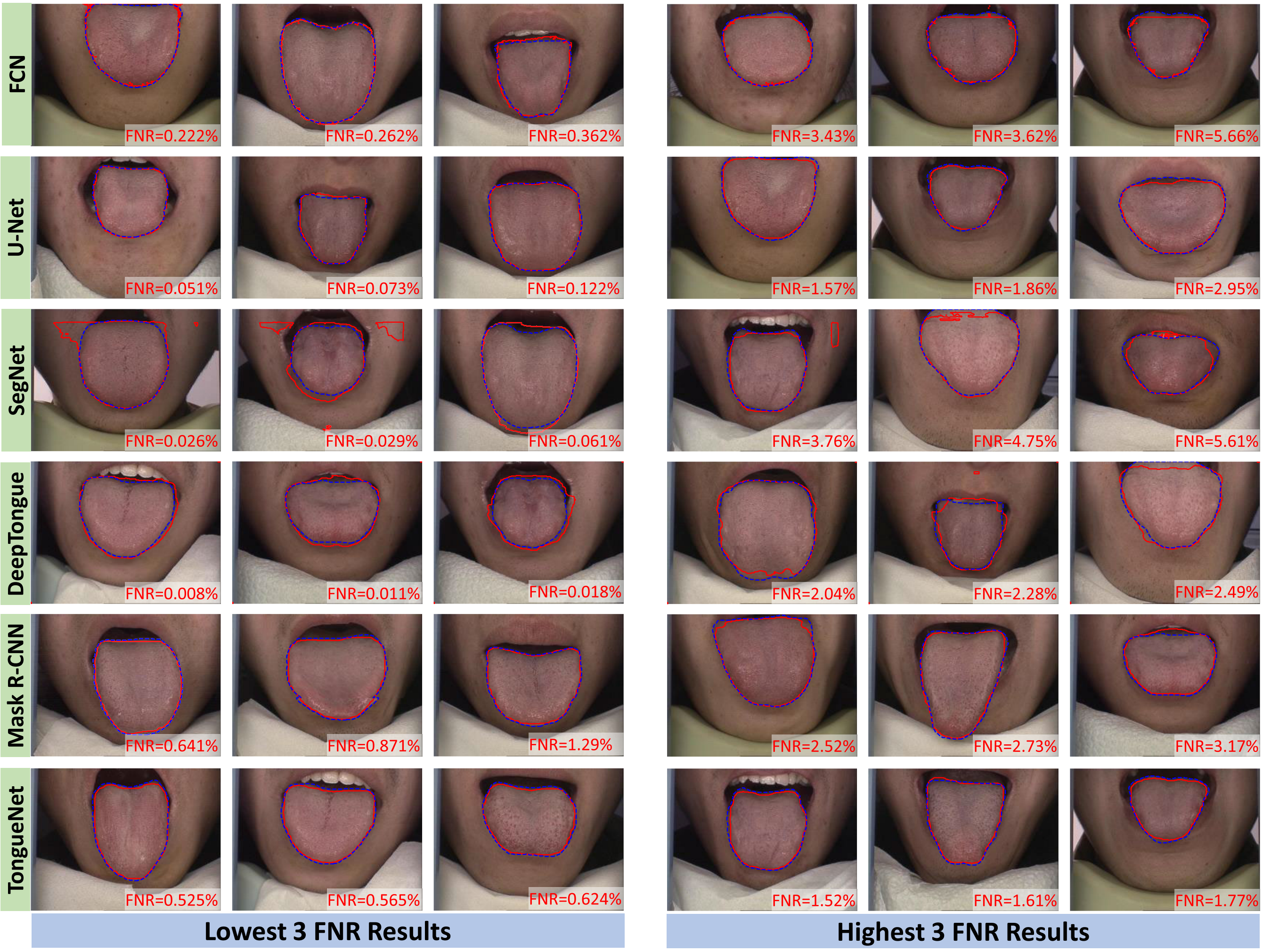
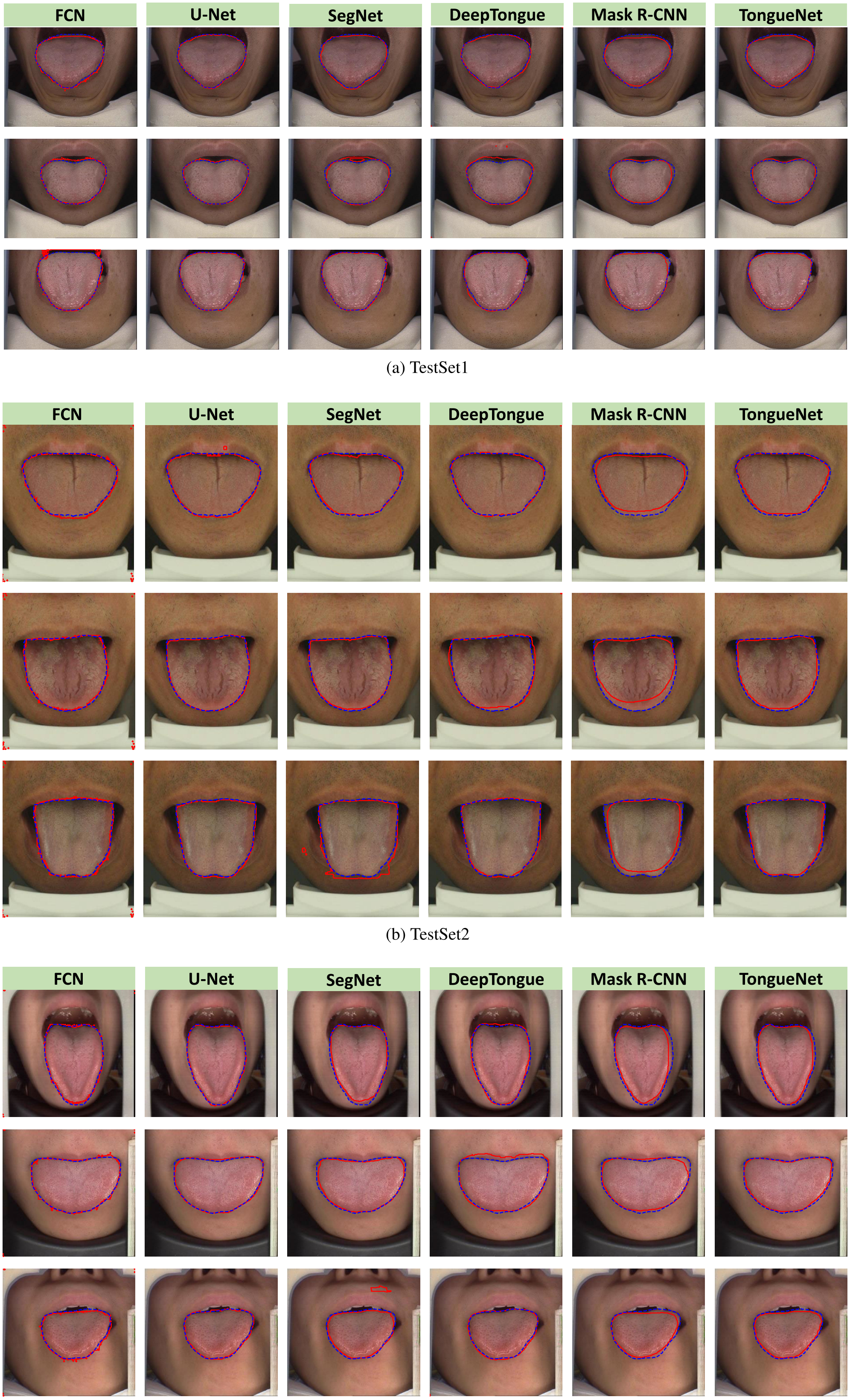


**FIGURE 9. The lowest 3 FNR and highest 3 FNR results of different methods on TestSet1. Red solid line indicates the prediction and Blue dashed line means the ground truth.**

tongue body, which results in a lower FNR but a much higher FPR value. As shown in Fig. 9, compared with other baselines which have relatively lower average FNR, TongueNet performs more stable and achieves better closeness between the prediction and the boundary of tongue body. Moreover, compared with Mask R-CNN, a multi-task learning based



(a) TestSet1

(b) TestSet2

(c) TestSet3

**FIGURE 10. Visualization results from different methods experimented on three datasets. Red solid line indicates the prediction and Blue dashed line means the ground truth.**

method, TongueNet achieves stable improvement on all metrics especially for FNR score because of the down-sample of feature maps in the original feature pyramid network of Mask R-CNN. It demonstrates the effectiveness of the proposed context-aware feature extraction module and the Tversky index based loss function used for the issue of large class imbalance.

### 2) QUALITATIVE RESULTS

To demonstrate the effectiveness of the proposed method qualitatively, experimental results of different methods on three datasets are shown in Fig. 10. Three randomly selected samples from TestSet1, TestSet2 and TestSet3 are displayed for qualitative visualization in Fig. 10a, 10b and 10c, respectively. It can be seen that most methods are capable of segmenting the tongue bodies with simple background accurately, as demonstrated in Fig. 10c. Those tongue images are captured from an uniform illumination condition and with less disturbance of other materials such as lips, tongue texture and coating. However, with large variations of tongue appearance, e.g. tongue texture, shape, and coating of different patients, TongueNet achieves not only high segmentation accuracy, but also a more stable performance on all datasets. For example, as shown in Fig. 10a and Fig. 10b, for FCN and SegNet, those algorithms conduct segmentation directly on the whole image, which are easily misled by the complex materials such as lips and face. Different from those methods, the proposed TongueNet carries out the segmentation of tongue body only on the feature maps of ROIs localized by the RPN, as illustrated in Fig. 7, which narrows the scope of segmentation to alleviate the influence of hard samples like lips and face. Compared with Mask R-CNN, TongueNet achieves more accurate segmentation results which demonstrates the effectiveness of the proposed context-aware feature extraction module and the Tversky index based loss function used for training. Combining with the results shown in Table 1 and Fig. 8, it is obvious that the proposed TongueNet leads to the most accurate and robust performance compared with all other methods.

## V. CONCLUSION

In this paper, we study the problem of tongue localization and segmentation, which is vital in an automatic tongue diagnostic system in traditional Chinese medicine and challenging because of the issues of large variations of tongue appearance, data imbalance, and hard sample mining. To solve the above-mentioned issues, we propose a new end-to-end tongue localization and segmentation method, named TongueNet, which segments tongue in a pixel-to-pixel manner automatically based on deep convolution neural network. To extract more discriminative multi-scale features of tongue images, a novel feature pyramid network based on the designed context-aware residual blocks is proposed. Then, based on the extracted multi-scale feature maps, TongueNet is able to localize the ROI of tongue body rapidly, and then generates a precise segmentation mask of tongue body based on the extracted ROI, which does not require any preprocessing and is robust to shape deformation, color variations of tongue coating and different illumination conditions. Quantitative and qualitative comparisons on real-world datasets show that the proposed TongueNet achieves state-of-the-art performance for the segmentation of tongue body in terms of both robustness and accuracy.

## REFERENCES

[1] H. Z. Zhang, K. Q. Wang, D. Zhang, B. Pang, and B. Huang, "Computer aided tongue diagnosis system," in *Proc. IEEE Eng. Med. Biol. 27th Annu. Conf.*, Jan. 2006, pp. 6754–6757.

[2] N. Sharma and L. M. Aggarwal, "Automated medical image segmentation techniques," *J. Med. Phys.*, vol. 35, no. 1, pp. 3–14, Jan. 2010.

[3] J. Wu, Y. Zhang, and J. Bai, "Tongue area extraction in tongue diagnosis of traditional Chinese medicine," in *Proc. IEEE Eng. Med. Biol. 27th Annu. Conf.*, Jan. 2006, pp. 4955–4957.

[4] J. Ning, D. Zhang, C. Wu, and F. Yue, "Automatic tongue image segmentation based on gradient vector flow and region merging," *Neural Comput. Appl.*, vol. 21, no. 8, pp. 1819–1826, Nov. 2012.

[5] B. Pang, D. Zhang, and K. Wang, "The bi-elliptical deformable contour and its application to automated tongue segmentation in Chinese medicine," *IEEE Trans. Med. Imag.*, vol. 24, no. 8, pp. 946–956, Aug. 2005.

[6] M. Shi, G. Li, and F. Li, "C$^2$G$^2$FSnake: Automatic tongue image segmentation utilizing prior knowledge," *Sci. China Inf. Sci.*, vol. 56, no. 9, pp. 1–14, Sep. 2013.

[7] X. Zhai, H.-D. Lu, and L. Zhang, "Application of image segmentation technique in tongue diagnosis," in *Proc. Int. Forum Inf. Technol. Appl.*, vol. 2, May 2009, pp. 768–771.

[8] K. Wu and D. Zhang, "Robust tongue segmentation by fusing region-based and edge-based approaches," *Expert Syst. Appl.*, vol. 42, no. 21, pp. 8027–8038, Nov. 2015.

[9] Z. Li, Z. Yu, W. Liu, and Z. Zhang, "Tongue image segmentation via color decomposition and thresholding," in *Proc. 4th Int. Conf. Inf. Sci. Control Eng. (ICISCE)*, Jul. 2017, pp. 752–755.

[10] J. Li, B. Xu, X. Ban, P. Tai, and B. Ma, "A tongue image segmentation method based on enhanced HSV convolutional neural network," in *Proc. Int. Conf. Cooperat. Design, Visualizat. Eng.* Cham, Switzerland: Springer, 2017, pp. 252–260.

[11] P. Qu, H. Zhang, L. Zhuo, J. Zhang, and G. Chen, "Automatic tongue image segmentation for traditional Chinese medicine using deep neural network," in *Proc. Int. Conf. Intell. Comput.* Cham, Switzerland: Springer, 2017, pp. 247–259.

[12] B. Lin, J. Xie, C. Li, and Y. Qu, "Deeptongue: Tongue segmentation via resnet," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 1035–1039.

[13] P. O. Pinheiro, T.-Y. Lin, R. Collobert, and P. Dollár, "Learning to refine object segments," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 75–91.

[14] F. Yu, V. Koltun, and T. Funkhouser, "Dilated residual networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 472–480.

[15] W. Zuo, K. Wang, D. Zhang, and H. Zhang, "Combination of polar edge detection and active contour model for automated tongue segmentation," in *Proc. 3rd Int. Conf. Image Graph. (ICIG)*, Dec. 2004, pp. 270–273.

[16] X. Li, J. Li, and D. Wang, "Automatic tongue image segmentation based on histogram projection and matting," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Nov. 2014, pp. 76–81.

[17] Y. Zheng and C. Kambhamettu, "Learning based digital matting," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep./Oct. 2009, pp. 889–896.

[18] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Single-shot refinement neural network for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 4203–4212.

[19] H.-M. Yang, X.-Y. Zhang, F. Yin, and C.-L. Liu, "Robust classification with convolutional prototype learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 3474–3482.

[20] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3431–3440.

[21] H. Fan, F. Zhang, L. Xi, Z. Li, G. Liu, and Y. Xu, ''LeukocyteMask: An automated localization and segmentation method for leukocyte in blood smear images using deep neural networks,'' *J. Biophoton.*, vol. 12, Jul. 2019, Art. no. e201800488.

[22] V. Badrinarayanan, A. Kendall, and R. Cipolla, ''SegNet: A deep convolutional encoder-decoder architecture for image segmentation,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[23] Y. Xue, X. Li, P. Wu, J. Li, L. Wang, and W. Tong, ''Automated tongue segmentation in Chinese medicine based on deep learning,'' in *Proc. Int. Conf. Neural Inf. Process.* Cham, Switzerland: Springer, 2018, pp. 542–553.

[24] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, ''Rethinking atrous convolution for semantic image segmentation,'' 2017, *arXiv:1706.05587*. [Online]. Available: https://arxiv.org/abs/1706.05587

[25] K. He, X. Zhang, S. Ren, and J. Sun, ''Deep residual learning for image recognition,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

[26] R. Caruana, ''Multitask learning,'' *Mach. Learn.*, vol. 28, no. 1, pp. 41–75, 1997.

[27] A. Maurer, M. Pontil, and B. Romera-Paredes, ''The benefit of multitask representation learning,'' *J. Mach. Learn. Research*, vol. 17, no. 1, pp. 2853–2884, Jan. 2016.

[28] S.-A. Rebuffi, H. Bilen, and A. Vedaldi, ''Learning multiple visual domains with residual adapters,'' in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 506–516.

[29] K. He, G. Gkioxari, P. Dollár, and R. Girshick, ''Mask R-CNN,'' in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2961–2969.

[30] S. Liu, E. Johns, and A. J. Davison, ''End-to-end multi-task learning with attention,'' 2018, *arXiv:1803.10704*. [Online]. Available: https://arxiv.org/abs/1803.10704

[31] I. Misra, A. Shrivastava, A. Gupta, and M. Hebert, ''Cross-stitch networks for multi-task learning,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 3994–4003.

[32] I. Kokkinos, ''UberNet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6129–6138.

[33] K. Simonyan and A. Zisserman, ''Very deep convolutional networks for large-scale image recognition,'' 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556

[34] F. Yu and V. Koltun, ''Multi-scale context aggregation by dilated convolutions,'' in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2016, pp. 1–13.

[35] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, ''DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2017.

[36] A. Krizhevsky, I. Sutskever, and G. E. Hinton, ''Imagenet classification with deep convolutional neural networks,'' in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[37] T.-Y. Lin and P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, ''Feature pyramid networks for object detection,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2117–2125.

[38] S. Ren, K. He, R. Girshick, and J. Sun, ''Faster R-CNN: Towards real-time object detection with region proposal networks,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[39] R. Girshick, ''Fast R-CNN,'' in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1440–1448.

[40] A. Tversky, ''Features of similarity,'' *Psychol. Rev.*, vol. 84, no. 4, pp. 327–352, Jul. 1977.

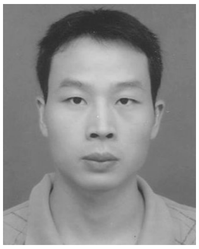
[41] S. R. Hashemi, S. S. M. Salehi, D. Erdogmus, S. P. Prabhu, S. K. Warfield, and A. Gholipour, ''Tversky as a loss function for highly unbalanced image segmentation using 3D fully convolutional deep networks,'' 2018, *arXiv:1706.05721*. [Online]. Available: https://arxiv.org/abs/1706.05721

[42] X. Wang and D. Zhang, ''An optimized tongue image color correction scheme,'' *IEEE Trans. Inf. Technol. Biomed.*, vol. 14, no. 6, pp. 1355–1364, Nov. 2010.

[43] D. P. Kingma and J. Ba, ''Adam: A method for stochastic optimization,'' in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–15.

[44] E. Shelhamer, J. Long, and T. Darrell, ''Fully convolutional networks for semantic segmentation,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.

[45] O. Ronneberger, P. Fischer, and T. Brox, ''U-Net: Convolutional networks for biomedical image segmentation,'' in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2015, pp. 234–241.



**CHANGEN ZHOU** received the B.S. degree from the School of Mathematics, Henan Normal University, Xinxiang, China, in 2004, and the M.S. degree in computer science and technology from Fuzhou University, Fuzhou, China, in 2007. He is currently a Lecturer with the College of Traditional Chinese Medicine, Fujian University of Traditional Chinese Medicine, Fuzhou. His current research interests include image processing and machine learning.

**HAOYI FAN** received the B.S. degree from the School of Mathematics, Zhengzhou University of Aeronautics, Zhengzhou, China, in 2015. He is currently pursuing the Ph.D. degree with the School of Computer Science and Technology, Harbin University of Science and Technology, Harbin, China. His current research interests include pattern recognition, information security, and machine learning.

**ZUOYONG LI** received the B.S. and M.S. degrees in computer science and technology from Fuzhou University, Fuzhou, China, in 2002 and 2006, respectively, and the Ph.D. degree from the School of Computer Science and Technology, Nanjing University of Science and Technology (NUST), Nanjing, China, in 2010. He is currently a Professor with the College of Computer and Control Engineering, Minjiang University, Fuzhou. He has published more than 60 articles in international/national journals. His current research interests include image processing, pattern recognition, and machine learning.