IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 29, 2020, accepted May 8, 2020, date of publication May 18, 2020, date of current version June 1, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2995202

TCMINet: Face Parsing for Traditional Chinese Medicine Inspection via a Hybrid Neural Network With Context Aggregation

XINLEI LI[1], DAWEI YANG[1], YAN WANG[1], WEI ZHANG[2], FUFENG LI[3], AND WENQIANG ZHANG[1,2]

[1]Academy for Engineering and Technology, Fudan University, Shanghai 200433, China
[2]Shanghai Key Laboratory of Intelligent Information Processing, Fudan University, Shanghai 200433, China
[3]Laboratory of TCM four Processing, Shanghai University of Traditional Chinese Medicine, Shanghai 201203, China

Corresponding author: Wenqiang Zhang (wqzhang@fudan.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 81774205, in part by the Special Fund of the Ministry of Education of China under Grant 2018A11005, in part by the Fudan University-Changchun Institute of Optics, Fine Mechanics and Physics (CIOMP) Joint Fund under Grant FC2019-005, and in part by the Jihua Laboratory under Grant Y80311W180.

ABSTRACT Facial medical analysis, including the inspection of the face and inner facial components, has always been a primary part of the diagnostic method in Traditional Chinese Medicine (TCM). The existing literature merely focus on detecting or segmenting single face organs such as tongue, eyes, or lips. In this paper, we make the first attempt to deal with multiple organs simultaneously and develop an end-to-end hybrid network with context aggregation (named TCMINet) to achieve face parsing for Traditional Chinese Medicine Inspection (TCMI). Additionally, we construct a new dataset named TCMID to overcome the lackness of accurate annotated data. In order to verify the generalization ability of TCMINet, we manually relabel images in two popular face parsing datasets referred to as LFW-PL⋆ and HELEN⋆ for test. The extensive ablation evaluations and experimental comparisons demonstrate that the proposed TCMINet outperforms state-of-the-art methods under various evaluation metrics. It runs at 267ms per face (512×512 image) on Nvidia Titan Xp GPU, being possible to be integrated into engineering solutions.

INDEX TERMS Traditional Chinese medicine inspection, semantic segmentation, face parsing, hybrid neural networks, context aggregation.

I. INTRODUCTION

Nowadays, Traditional Chinese Medicine (TCM) has become a global and essential diagnostic approach in the medical field [1]. In TCM, inspection is a critical diagnostic step to check the current state of patients with an observation of the expression, appearance, color, and abnormal changes of the body, face, and inner facial components (e.g., eyes, lips, tongue). The face and inner facial components are believed to reveal signs of various health conditions or even diseases of the internal body [2]. For instance, people with hepatitis and other liver issues may have a face or eyes with a yellow tone [3]. The tongue of HIV-infected patients may be swollen, and tooth marked [4]. Moreover, the lip color of a person is considered as a symptom to reflect the physical conditions of organs in the body [5].

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

Generally, as preprocessing, the first step of most computer vison-aided facial medical analysis techniques consists in detecting or segmenting face and facial components from face images. However, the existing literature merely focus on detecting or segmenting single face organs [6]–[8]. As a special case of face parsing, face parsing for TCMI amounts to labeling each pixel with the left eye, right eye, lips, tongue, face, and background, following the principles of TCM holistic view [9], [10]. Inevitably, some challenging problems hiding behind this task are as follows. First, the patient opens the mouth wide with the tongue sticking out, and the lower lip is partially (or totally) blocked by the tongue. Second, the tongue color gamut is highly overlapping with lip (face) color gamut. Third, in addition to the face and target facial components, obtained face images contain many other non-target components, such as hair, beard, teeth, and the inner tissue of the mouth. And fourth, the patient's tongue color, facial expression, skin gloss, and other conditions are more

varied than healthy people. There are abundant pathological details on the surface of the patient's tongue, such as tongue crack, red point, tooth marks and etc. These details are often with only several-pixel size, which makes parsing more difficult.

Existing face parsing literature [11]–[13] have illustrated significant advantages by focusing on individual regions of interest (ROIs) for inner facial components. However, these methods [11]–[15] mainly focus on segmenting hair, eyebrows, and other facial components that are rarely relevant to TCMI, rather than segmenting the tongue that is essential for TCMI applications. Face parsing for TCMI is indeed a new challenging task, and too little work has been devoted to this area. Accordingly, proposing a new hybrid architecture that follows the TCM diagnostic principles is of great need. Furthermore, in order to parse face images robustly, effective contextual modeling [16]–[19] is more demanding. Inspired by these methods, we propose a novel TCMINet to estimate masks for each face and each inner facial component separately, which is shown as Fig.1. Specifically, we first construct the Inspection Feature Extraction (IFE) module to complete efficient dense feature extraction with fast computation. Then the hierarchical Facial Inner Components Segmentation (FICS) structure and Face Segmentation (FS) structure are used to process the inner facial components (left eye, right eye, tongue, lips) and the face, respectively. Moreover, we employ context aggregation modules ($C1^{\uparrow\downarrow\leftarrow\rightarrow}$, $C2^{\searrow\nwarrow\nearrow\swarrow}$) to smooth the label prediction map as well as to refine boundary localization for inner facial components and the face. The symbol "arrow" indicates "the sweeping direction or propagation direction". For inner facial components, we employ an effective context aggregation module $C1^{\uparrow\downarrow\leftarrow\rightarrow}$ [21], [22], which uses four recurrent neural networks to sweep both vertically and horizontally along both directions across the image to incorporate the global context. For the face, we employ an efficient context aggregation module ($C2^{\searrow\nwarrow\nearrow\swarrow}$) [16], [17], which models semantic contextual dependencies of local representations with four context propagation directions (southeast, southwest, northwest, and northeast.

In addition to a high performance network, a good dataset with high-quality and well-labeled images is also a crucial component. There are only a few face parsing datasets, such as the LFW-PL [20] and HELEN [14]. Moreover, most of the images of [14], [20] are not suitable for this task. To mitigate this problem, we construct a face parsing dataset named TCMID, which contains 1500 face images captured by professional imaging devices under certain conditions (in a dark chest, not in open-air). In TCMID, each image is provided with accurate annotation of a 6-category (left eye, right eye, lips, tongue, face, and background) pixel-level label map. The contributions of this paper are summarized as follows:

1 We build a face parsing dataset (TCMID) and benchmark for training and test. To the best of our knowledge, it is the first face parsing dataset for TCMI. Furthermore, we manually relabel some images of HELEN and LFW-PL datasets named LFW-PL⋆ and HELEN⋆ for test. Datasets are available at: `https://github.com/FDUXilly/TCMID-face-image-dataset`.
2 We propose an effective hybrid architecture to address the problem of pixel-wise face parsing for TCMI. We introduce the context aggregation modules ($C1^{\uparrow\downarrow\leftarrow\rightarrow}$, $C2^{\searrow\nwarrow\nearrow\swarrow}$) that can significantly smooth the label prediction map as well as refine boundary localization for inner facial components and the face, respectively.
3 Our network surpasses previous state-of-the-art results on LFW-PL⋆, HELEN⋆, and TCMID datasets. Besides, ablation studies and exploratory experiments on TCMID are carried out to evaluate the hybrid network structure and important modules of our network. It runs at 267ms per face image (512×512) on a GPU, being possible to be integrated into engineering solutions.

II. RELATED WORK

A. TCMI - FACIAL MEDICAL ANALYSIS

Facial medical analysis is a non-contact, non-invasive diagnostic method of TCM [23]. Basically, the first task in the computer-aided facial medical analysis is to detect and segment the facial components from face images. In [2], five facial regions (Forehead, Left cheek, Right cheek, Nose, Jaw) are segmented using skin detection, facial normalization, and horizontal position of the mouth, nostril, and eyebrow location. Hu *et al.* [6] adopt Gaussian Mixture Model (GMM) in lip segmentation. Li *et al.* [7] propose an end-to-end iterative tongue image matting network. Rot *et al.* [8] present a deep multi-class eye segmentation model build upon the SegNet architecture. As mentioned before, these methods only take separate face organs into account, resulting in inaccurate and biased diagnostic results. In this paper, we propose a hybrid architecture that can simultaneously detect and segment multiple facial components based on the principles of TCM holistic view.

B. SEMANTIC SEGMENTATION

Semantic segmentation is more and more being of interest for computer vision researchers. FCN [24] is a baseline for generic images which employs full convolution on the entire image to extract feature. Mask R-CNN [25] further advances the cutting edge of semantic segmentation through extending Faster R-CNN [26] and integrating RoIAlign. Mask Scoring R-CNN [27] extends Mask R-CNN with MaskIoU Head and achieves a new state-of-the-art result. However, directly applying these methods for face parsing may fail to model the complex yet varying spatial layout across face components, leading to unsatisfactory results.

1) CONTEXT AGGREGATION

One major group of works focus on context aggregation dependencies of local regions in the CRF framework [18],

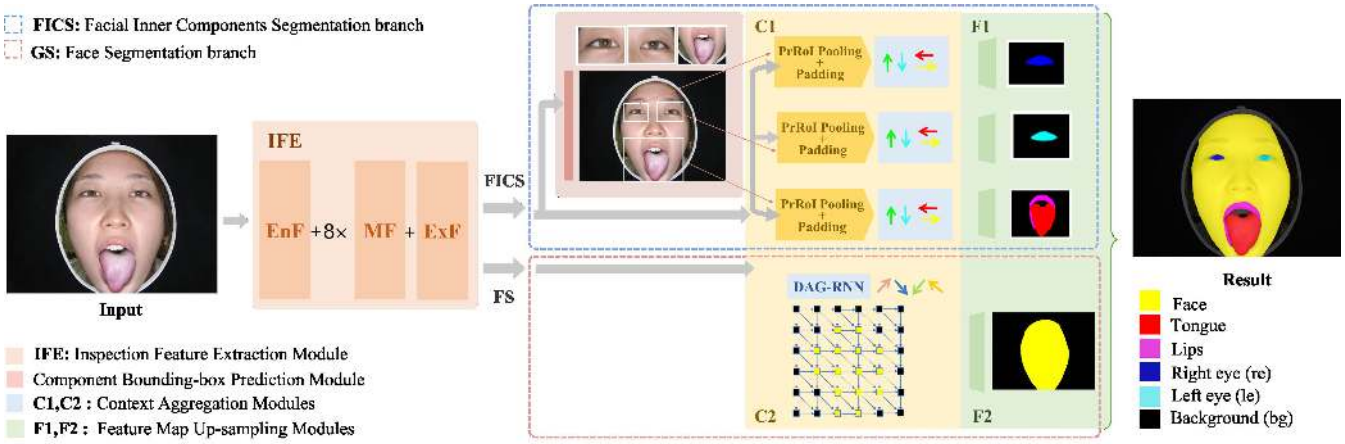

FIGURE 1. **Proposed network structure. Our basic segmentation network essentially consists of two branches (FICS, FS) and four functional modules: Inspection Feature Extraction (IFE), Component Bounding-box (Bbox) Prediction, Context Aggregation (C1$^{\uparrow\downarrow\leftarrow\rightarrow}$, C2$^{\searrow\nwarrow\nearrow\swarrow}$), and feature map up-sampling (F1,F2).**

[19]. Another group of works introduce sub-networks that can aggregate context inherently [28].

C. FACE PARSING

Face parsing, aiming to assign pixel-level semantic labels for face images, has attracted much attention due to its wide application potentials, such as: facial beautification [29], face image synthesis [30].

1) METHODS

Most existing approaches for face parsing can be categorized into two branches: global-based methods [31], [32] and local-based (hybrid) methods [11]–[13], [15]. Global-based methods predict semantic labels over the whole input image. Wei *et al.* [31] design automatically regulating receptive fields in a deep image parsing network. Zhou *et al.* [32] propose a network that employs super-pixel information and the CRF model jointly. Nevertheless, the accuracy of these kinds of methods is limited due to the lack of focusing on each individual part (see Table 10). In contrast, local-based methods train separated models for various facial components. Zhou *et al.* [11] design an interlinked CNN-based architecture which predicts pixel labels after facial localization. Liu *et al.* [12] propose a network that combines hierarchical representations learned by a CNN, and label propagations achieved by a spatially variant RNN. Lin *et al.* [13] propose a novel network combined with RoI Tanh-warping for face parsing. All of these approaches focus on general facial parsing tasks but ignore some relevant facial components that are essential for TCMI applications.

2) DATASETS

Although many face related fields have been well studied for many years, the existing datasets for face parsing are still severely limited. This is mainly because pixel-level annotation is a time-consuming work. The most commonly used public datasets for face parsing methods are LFW-PL [20] and HELEN [14]. LFW-PL dataset contains 2,927 face images. All the images are manually assigned to one of the hair/skin/background categories. The HELEN dataset contains 2,330 face images with manually labeled facial components including eyes, eyebrows, nose, inside mouth, lips, etc. However, for both datasets, the tongue in each image is not annotated.

D. RECURRENT NEURAL NETWORK

RNNs have been shown to be effective for modeling short and long term dependencies in sequential data. For images, we can apply 1-D RNN to multiple dimensions [21], [33] or multi-dimensional RNN (MDRNN) [34], [35] such that each neural node can receive informations from multiple directions. Visin *et al.* [21] adopt ReNet, which is a stacked of 1D-RNN to perform image classification. Based on ReNet model, an architecture for semantic segmentation called ReSeg [22] has been proposed. Both of them observe promising performance gains after incorporating RNNs.

III. THE PROPOSED TCMINET

We use a hybrid solution to estimate masks for face and inner facial components simultaneously. Given a cropped face image I, which contains only a single face in the center of the image, the Inspection Feature Extraction (IFE) module is deployed to capture dense feature maps F, which are later shared by Facial Inner Components Segmentation (FICS) and Face Segmentation (FS) branches. In FICS branch, for each inner facial component $\{C_i\}_{i=1}^{N} = \{$left eye, right eye, lips, tongue$\}$ where N is the number of individual component, the local bounding-box (bbox) $\{R_i\}_{i=1}^{N-1}$ of each component C_i is predicted from F. The features of each component within their bbox are mapped to a fixed resolution through PrRoI Pooling [36]. Next, C1$^{\uparrow\downarrow\leftarrow\rightarrow}$ is adopted to model global contexts and reduce computational cost. At the end of the FICS branch, the segmentation masks $\{M_i\}_{i=1}^{N-1}$ for each component are predicted individually. Meanwhile, in the FS branch, C2 $^{\searrow\nwarrow\nearrow\swarrow}$ is designed to link pixel-level and local information of F. Same as the FICS branch, pixel-wise

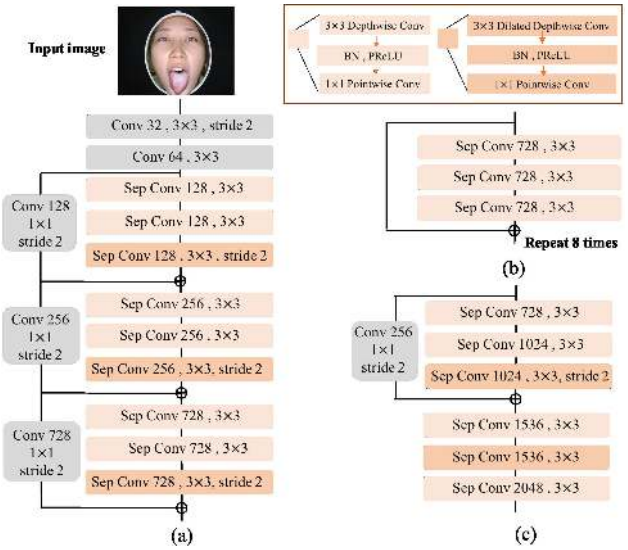

FIGURE 2. The architecture of the inspection feature extraction module IFE. (a) the Entry Flow (EnF), (b) the Middle Flow (MF), (c) the Exit Flow (ExF). The input image *I* first goes through the EnF, then through the MF which is repeated eight times, and finally through the ExF.

segmentation mask M_{face} is predicted in the end. Finally, we gather all segmentation masks and form the face parsing result as M.

As illustrated in Fig.1, we introduce the whole network with four temporal-consecutive functional modules: IFE, Component Bounding-box Prediction, Context Aggregation, and Feature Map Up-sampling. Next, we introduce each module in detail.

A. IFE MODUEL

The Xception model [37]–[39] has shown promising performance with fast computation. We work in the same direction to modify the Xception model for the task of face parsing. As illustrated in Fig. 2, max-pooling operations are replaced by depthwise separable convolutions, which allows efficient dense feature extraction on any arbitrary resolution. We use PReLU [40] as the non-linearity rather than ReLU since it allows negative responses that in turn improves the network performance (see Table 5). Furthermore, all of 3×3 depthwise convolution layers and 3×3 dilated depthwise convolution layers are followed by a BN and a PReLU activation.

B. FACIAL INNER COMPONENTS SEGMENTATION (FICS) BRANCH

1) COMPONENT BOUNDING-BOX PREDICTION MODULE

The semantic label of every inner facial component is explicitly defined in our work (e.g., left/right eye). Here we explicitly regress the area of each inner facial component instead of detecting them individually like in a Mask R-CNN-fashion [25], [27]. The prediction module consists of two convolutional layers followed by a global average pooling and a fully connected layer. It avoids ambiguities in components and reduces computation cost. The component prediction module locates bounding-boxes of the N inner facial components: $\{R_i\}_{i=1}^{N-1}$. The annotated ground truth bounding-boxes are denoted as $\{R_i^g\}_{i=1}^{N-1}$. We adopt the L_1 loss for the bounding-box regression. The regression loss L_{reg} is defined as:

$$L_{reg} = \frac{1}{N-1}\sum_{i}^{N-1} \parallel R_i - R_i^g \parallel_1 \tag{1}$$

It stands to reason that the low accuracy of the regressed bounding-boxes usually leads to the poor performance of the segmentation. Through experiments, we observe that some part of the targets may fall outside the bounding-boxes, especially for lips. To mitigate this problem, we add paddings outside the bounding-boxes to solve the problem. Specifically, regressed bounding-boxes are padded by 20% the feature map size for lips and 10% for other components. The optimized bounding-boxes yield good hints for predicting high accuracy masks (see Table 6).

2) CONTEXT AGGREGATION MODULE (C1$^{\uparrow\downarrow\leftrightarrow}$)

In order to parse face images robustly, effective contextual modeling is more demanding. For inner facial components $\{C_i\}_{i=1}^{N} = \{\text{left eye, right eye, lips, tongue}\}$, we first use PrRoI Pooling [36] to map the features of each component to a fixed resolution individually. We feed the resulting feature maps E_i into recurrent layers for fine-tuning. As depicted in Fig.1, each recurrent layer is composed by four RNNs. Specifically, we take a feature map E_i of elements $e \in \mathbb{R}^{A\times B\times C}$, where A, B and C are respectively the height, width and number of channels and we split it into $K \times L$ patches $p_{k,l} \in \mathbb{R}^{A_p\times B_p\times C}$. First, we sweep the image vertically with two RNNs ($o^{\downarrow}$ and $o^{\uparrow}$). Each RNN reads the next non-overlapping patch $p_{k,l}$ based on its previous state, emits a projection $q_{k,l}^{\downarrow}$ (or $q_{k,l}^{\uparrow}$) and updates its state $r_{k-1,l}^{\downarrow}$ (or $r_{k+1,l}^{\uparrow}$):

$$\begin{aligned} q_{k,l}^{\downarrow} &= o^{\downarrow}\left(r_{k-1,l}^{\downarrow}, p_{k,l}\right), \quad \text{for } k = 1, \cdots, K \\ q_{k,l}^{\uparrow} &= o^{\uparrow}\left(r_{k+1,l}^{\uparrow}, p_{k,l}\right), \quad \text{for } k = K, \cdots, 1 \end{aligned} \tag{2}$$

We concatenate projections $q_{k,l}^{\downarrow}$ and $q_{k,l}^{\uparrow}$ to obtain feature map $Q^{\updownarrow}$. Then we sweep over each of its rows with two RNNs ($o^{\leftarrow}$ and $o^{\rightarrow}$). With a similar but specular procedure as the one described before, we obtain a concatenated feature map $Q^{\leftrightarrow}$. Each element $q_{k,l}^{\leftrightarrow}$ represents the features of patches $p_{k,l}$ with contextual information from E_i. To sum up, the context aggregation module sweeps over feature maps E_i horizontally and vertically, and providing relevant global information.

3) FEATURE MAP UP-SAMPLING MODULE (F1)

All component's feature map up-sampling modules share the same network architecture but have independent weights. Each component segmentation module is built with two 3×3 convolutions each followed by one bilinear up-sampling. For the obtained $N - 1$ bounding-boxes, $N - 1$ light and parallel feature map up-sampling modules are used to predict

the masks for each inner facial component. We use the pixel-wise cross-entropy to measure the component segmentation accuracy. The segmentation loss L_{seg1} is defined as the averaged cross-entropy among all the segmentation networks:

$$L_{seg1} = \frac{1}{N-1}\sum_{i=1}^{N-1} CrossEntropy(M_i, M_i^g) \tag{3}$$

C. FACE SEGMENTATION (FS) BRANCH

1) CONTEXT AGGREGATION MODULE (C2↘↖↗↙)

Different from chain-structured sequential data, the connectivity structure of image units are beyond chain. The graphical representations (e.g., UCGs) that respect the 2-D neighborhood system are more plausible solutions for spatial arrangement of image units. However, due to the loopy structure of UCGs, RNNs can't be directly applied to UCG-structured images. We decompose the UCG $\mathcal{U}$ to a set of complimentary DAGs: $\mathcal{U} = \sum \mathcal{D}_d$. As exemplified in Fig.3, we use the four context propagation directions ($\mathcal{D}_1$ ↘, $\mathcal{D}_2$ ↖, $\mathcal{D}_3$ ↗ and $\mathcal{D}_4$ ↙) to decompose $\mathcal{U}$. The topology of feature maps F is represented as DAG $\mathcal{D} = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V} = \{v_i\}_{i=1:N}$ is the vertex set and $\mathcal{E} = \{e_{ij}\}$ is the arc set. The topology of the hidden layer h_d follows the same topology as $\mathcal{D}$. Therefore, a forward propagation sequence can be generated by traversing $\mathcal{D}$. These operations can be mathematically expressed as follows:

$$\mathbf{h}_d^{(v_i)} = g\left(M_d\mathbf{x}^{(v_i)} + \sum_{v_j\in\mathcal{P}_{\mathcal{D}_d}(v_i)} W_d\mathbf{h}_d^{(v_j)} + b_d\right) \tag{4}$$

$$\mathbf{o}^{(v_i)} = k\left(\sum_{\mathcal{D}_d} V_d\mathbf{h}_d^{(v_i)} + c\right) \tag{5}$$

where M_d, V_d, and W_d are weight matrices, and b_d is bias vector. Here, x^{v_i}, h^{v_i}, and o^{v_i} are the representations of input, hidden and output layers located at v_i, respectively. $\mathcal{P}_{\mathcal{D}_d}(v_i)$ is the direct predecessor set of vertex v_i in $\mathcal{D}_d$. g and k are composition functions. We place C2 ↘↖↗↙ on top of the IFE module to capture the rich contextual dependencies over image regions.

2) FEATURE MAP UP-SAMPLING MODULE (F2)

For the face, we perform several convolutions and up-sampling operations to generate the mask M_{face} (M_{face}^g is the groundtruth). We also use the cross-entropy loss to constrain the segmentation accuracy. The segmentation loss L_{seg2} is defined as:

$$L_{seg2} = CrossEntropy(M_{face}, M_{face}^g) \tag{6}$$

Finally, all the resulting segmentation masks are gathered. We form the final face parsing result, denoted as M. Since the component segmentation relies on a good component bounding-box regression, we divide the training process into two steps. In the first step, we only train the IFE module and the component bounding-box prediction module for good component regression accuracy. Here, only L_{reg} is used for training. In the second step, we perform joint training by updating all parameters with L_{reg}, L_{seg1}, and L_{seg2} together.

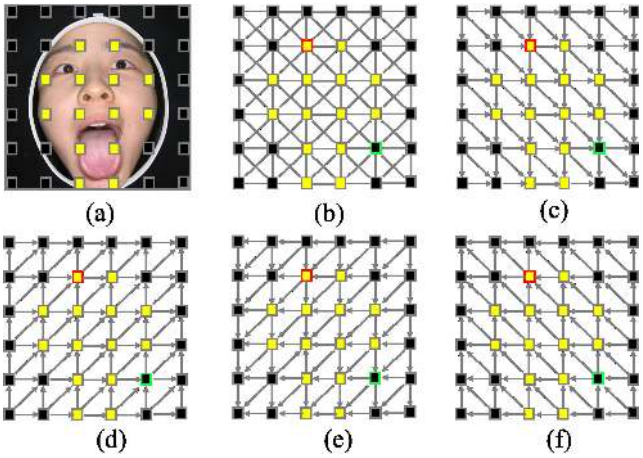

FIGURE 3. Illustration of context aggregation. (a): Feature tensor for a face image, and each square denotes one feature vector in the feature tensor. (b)-(f): The decomposition of $\mathcal{U}$ to four complimentary DAGs: {$\mathcal{D}_1$ ↘, $\mathcal{D}_2$ ↖, $\mathcal{D}_3$ ↗ and $\mathcal{D}_4$ ↙}. Note that any vertex pair (v_i, v_j) can be mutually communicable in DAGs. For example, local information of the red vertex can be routed to green vertex via $\mathcal{D}_1$, and green vertex can be routed to red vertex via $\mathcal{D}_4$.

IV. DATASETS

The dataset with diverse images and well-labeled masks is an important reason for the continuous improvement of face parsing algorithms, especially for deep learning-based technologies. To the best of our knowledge, there are only a few public face parsing datasets, such as the LFW-PL [20] and HELEN [14], where the hair area is considered as an essential semantic category for parsing. Especially, images in both datasets are taken in a random environment, and the tongue in each image is not annotated. The lack of accurate annotated datasets becomes a major obstacle in the progress of face parsing for TCMI. To fill the gap, we construct a novel dataset named TCMID, in which the tongue is regarded as one of the most critical semantic categories.

A. DATA COLLECTION

We collect 1500 face images in JPG format. The facial image acquisition system is the same as [5]. Table 1 shows the composition of our dataset. Besides, each image is provided with accurate annotation of a 6-category (face, left eye, right eye, lips, tongue, and background) pixel-level label map subjectively labeled by TCM practitioners. These images are split into the training and test sets with 1100 and 400 images, respectively.

B. IMAGE DIVERSITY

Face images in our dataset display large variations in (foreground) facial complexion, lip color, eye state, etc (see Table 2). As demonstrated in Fig.4, the patient's tongue (substance and coating) color, facial gloss, and other conditions

TABLE 1. Composition of the TCMID dataset.

Total	1500 subjects, 1500 face images, 1500 label maps			
Gender	Male		Female	
	708		792	
Ethnic group (of China)	Han	Miao	Zhuang	Hui
	526	329	296	349
Age group	<18	19∼37	38∼56	>57
	147	519	557	277
Health status	Healthy		Sub-healthy or unhealthy	
	374		1126	

TABLE 2. List of image attributes and the corresponding description.

Type	ID	Description	Type	ID	Description
Tongue substance Color	DR	Deep Red.	**Lips Color**	DER	Deep Red.
	LR	Light Red.		LIR	Light Red.
	P	Purple.		PU	Purple.
	R	Red.		RED	Red.
Tongue Coating Color	W	White.	**Eyes state**	O	Open.
	Y	Yellow.		HO	Half Open.
	G	Gray.		CL	Closed.
Mouth state	OP	Open.	**Face Gloss**	YES	Yes.
	HOP	Half Open.		L	Little.
	CLO	Closed.		N	NO.
Non-target Interference	H	Hair.	**Face Color**	WH	White.
	T	Teeth.		YE	Yellow.
	S	Saliva.		BL	Black.
	B	Bread.		RE	Red.
	I	Inner Tissues of the Mouth.		C	Cyan.
Rotation	RO	Roll			
	PI	Pitch			
	YA	Yaw			

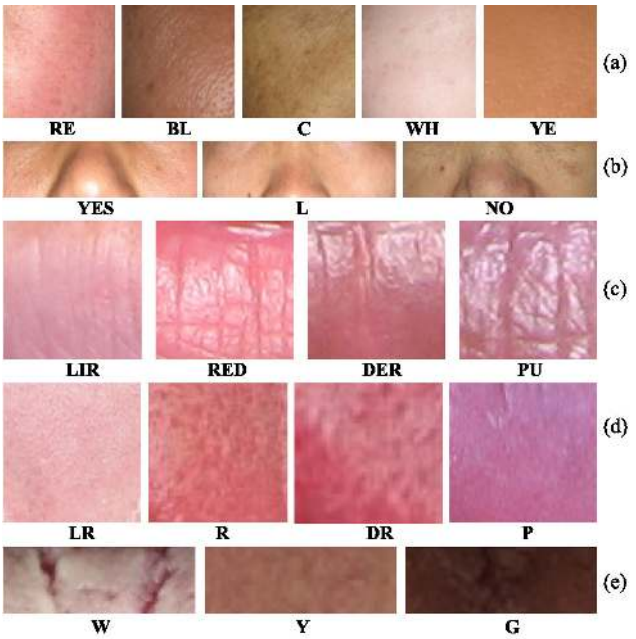

FIGURE 4. Typical complexion (color and gloss) squares with each label on the bottom. (a) Five typical facial colors. (b) Three facial gloss degrees. (c) Five typical lip colors. (d) Five typical tongue substance colors. (e) Three typical tongue coating colors.

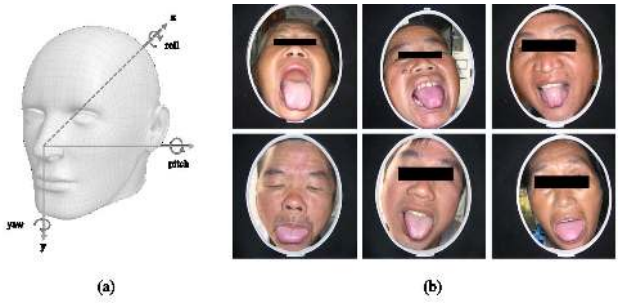

FIGURE 5. Head rotation. (a) The pose of the patient's head is described in the form of the three rotation angles yaw, pitch, and roll. (b) Sample images.

are greatly varied. The tongue substance color is usually reddish colors, and the tongue coating color is normally white, gray, or yellow. The tongue color gamut is highly overlapping with lip (face) color gamut. Fig.5 shows facial images with different head poses (rotation). As demonstrated in Fig.6.(c), in addition to the face and target inner facial components, typical face images inevitably contain many non-target components, such as hair, beard, teeth, and the inner tissue of the mouth. The different states (open, half-open, closed) of the eyes and mouth are shown in Fig.6.(a) and Fig.6.(b), respectively. Moreover, the patient opens the mouth wide with the tongue sticking out, and the lower lip is partially (or totally) blocked by the tongue. We include such large variations in TCMID to make our model more robust to challenging inputs.

C. DATA AUGMENTATION

While increasing the number of training images can enhance the performance of the model, we augment data by: (1) Geometric transformation. We exploit different rotating, resizing, and flipping to increase the number of training images. Four rotation angles $\{-45°, -20°, 20°, 45°\}$, four scales $\{0.5,0.8,1.2,1.5\}$ (around the center of the cropped face) are used. Horizontal flipping of images with probability 0.5 is used. (2) Gamma adjustment. We apply four different Gamma transforms to increase color variation. The Gamma values are $\{0.5,0.8,1.2,1.5\}$. (3) Background replacement. We first utilize a image matting network [42] to get the foreground (face) region. Then we randomly replace the background with non-face images [43], [44] or pure colors (e.g., deep red, light red, purple, red, white, yellow, gray).

D. OTHER DATASETS

Furthermore, we manually relabel some images of HELEN and LFW-PL datasets as challenge cases for test. It is worth noting that only a small number of face images in these two datasets conform to the TCMI face image-standard: the patient opens the mouth wide with the tongue sticking out. We selected face images that meet the standard in HELEN

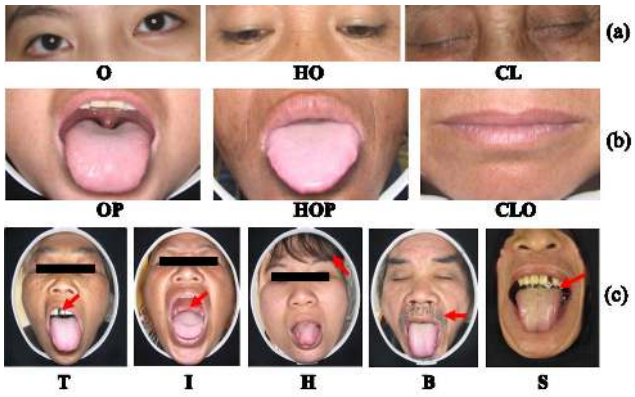

FIGURE 6. Typical face image samples with diverse facial component states or non-target facial component interference. (a) Eyes states (open, half-open, closed). (b) Mouth states (open, half-open, closed). (c) Non-target facial component interference. The Interference is indicated by red arrows.

TABLE 3. Image datasets of face parsing for TCMI.

Datasets	Purpose	Environment	Resolution	Samples	Labels
TCMID	Train/Test	Standard	512×512	1500	6
HELEN⋆	Test	Random	400×400	105	6
LFW-PL⋆	Test	Random	250×250	59	6

and LFW-PL datasets to build new test datasets. We refer to the relabeled test datasets as HELEN⋆ and LFW-PL⋆. LFW-PL⋆ dataset has 59 images for test. HELEN⋆ dataset has 105 images for test(see Table 3).

V. EXPERIMENTS

In this section, ablation studies and exploratory experiments on TCMID are carried out to discuss the hybrid network structure and several important modules of the proposed architecture. Then we test our network on the HELEN⋆, LFW-PL⋆, and TCMID datasets. Experimental results show that our model achieves the best results over other state-of-the-art methods on three datasets.

A. PERFORMANCE EVALUATION METRICS

Similar to [11]–[14], we use F-measure for each class as basic evaluation metrics. Besides, we quantitatively evaluate and compare our model with existing face parsing methods and semantic segmentation methods using evaluation metrics: Accuracy, Precision, Recall, F-measure, and their corresponding standard deviations. The metrics are defined as follows:

$$Accuracy = \frac{TP + FN}{TP + FN + FP + TN} \tag{7}$$

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

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

TABLE 4. Comparison with global-based and local-based methods on TCMID.

	Methods	le	re	lips	tongue	face	bg	mean	Acc.
Global -based	Smith et al. [14]	80.73	–	77.86	–	83.56	–	80.72	82.73
	Zhou et al. [11]	87.79	–	84.76	–	–	–	86.28	87.78
	Wei et al. [31]	85.19	–	–	–	91.65	–	88.42	89.13
Local -based	Liu et al. [12]	91.67	–	88.64	–	95.74	–	92.02	94.05
	Lin et al. [13]	92.89	92.93	90.63	–	96.37	–	93.21	95.55
	Ours	**95.91**	**95.89**	**93.46**	**96.98**	**97.95**	**98.27**	**96.04**	**98.33**

$$F = 2 \times \frac{precision \cdot recall}{precision + recall} \tag{10}$$

where TP denotes the number of true positive pixels, TN denotes the number of true negative pixels, FP stands for the number of false positive pixels, and FN represents the number of false negative pixels. The F-measure is the harmonic mean of Precision and Recall.

B. ABLATION STUDY

We quantitatively evaluate and compare our ablation models using Accuracy and F-measure metrics on TCMID dataset. The performances are reported in the form of F-measure for each class, mean F-measure over the five foreground categories (le, re, lips, tongue, face), and average Accuracy (Acc.). Herein, le is short for the left eye, re is short for the right eye, bg is short for the background.

1) IMPORTANCE OF THE HYBRID NETWORK STRUCTURE

We use a hybrid (local-based) strategy to train separated branches for face and detailed inner facial components. As explained in Section II.C, global-based methods directly predict the per-pixel semantic label over the whole face image. Table4 illustrates the advantages of hybrid structures over global-based structures in F-measure and average Accuracy. Experimentally, the accuracy of global-based methods [11], [14], [31] is limited due to the lack of focusing on each individual part.

2) IMPORTANCE OF THE IFE MODUEL

We found that the IFE module significantly gets better performance (compared with [24], [37], [38], [41]). As shown in Table 5, using IFE moduel as feature extractor achieves the state-of-the-art performance in terms of the highest Acc. on TCMID dataset. Meanwhile, adopting PReLU [40] as the non-linearity brings (le:0.23, re:0.25, lips:0.29, tongue:0.26, face:0.22, bg:0.25, mean:0.25) F-measure score improvement than using ReLU.

3) IMPORTANCE OF THE COMPONENT BOUNDING-BOX (BBOX) PREDICTION MODULE

As shown in Table 6, the component bounding-box prediction brings significant improvement (le:8.22, re:8.16, lips:14.23, tongue:12.25, face:4.74, bg:3.38, mean:9.52, Acc.:5.22) score for inner facial components segmentation, especially

TABLE 5. Comparison with different feature extractors on TCMID.

Methods	le	re	lips	tongue	face	bg	mean	Acc.
Ours(VGG16 [24])	92.09	92.11	90.41	95.88	95.56	96.72	93.21	95.83
Ours(ResNet [41])	93.71	93.79	91.23	94.73	96.61	96.89	94.01	97.04
Ours(Xception [37])	94.00	94.05	91.79	95.02	97.25	97.26	94.42	98.62
Ours([38])	94.12	94.15	91.94	95.13	97.43	97.89	94.55	97.83
Ours(IFE+ReLU)	95.68	95.64	93.17	96.72	97.73	98.02	95.79	98.11
Ours(IFE+PReLU)	**95.91**	**95.89**	**93.46**	**96.98**	**97.95**	**98.27**	**96.04**	**98.33**

TABLE 6. Importance of the component bbox prediction module.

Methods	le	re	lips	tongue	face	bg	mean	Acc.
FCN [24]	83.12	–	76.51	78.38	84.36	86.92	80.59	82.74
Ours(w/o bbox)	87.69	87.73	79.23	84.73	93.21	94.89	86.52	93.11
Ours(w/o padding)	95.18	95.21	92.01	95.87	96.83	96.77	95.02	97.03
Ours	**95.91**	**95.89**	**93.46**	**96.98**	**97.95**	**98.27**	**96.04**	**98.33**

TABLE 7. Comparison with different Up-sampling methods on TCMID.

Methods	le	re	lips	tongue	face	bg	mean	Acc.
Ours(weightsharing)	95.31	95.33	92.97	96.28	97.86	98.03	95.55	98.05
Ours(separated)	**95.91**	**95.89**	**93.46**	**96.98**	**97.95**	**98.27**	**96.04**	**98.33**

for 'lips' and 'tongue'. In addition, padding (20% the feature map size for lips and 10% for other components) the regressed bounding-boxes brings another (le:0.73, re:0.68, lips:1.45, tongue:1.11, face:1.12, bg:1.50, mean:1.02, Acc.:1.3) score improvement.

4) IMPORTANCE OF SEPARATED SEGMENTATION MODULES

Different from [25], [27], our segmentation modules do not share weights. The importance of separated weights is verified by the results from Table 7. Adopting "separated weights" strategy brings (le:0.60, re:0.56, lips:0.79, tongue:0.70, face:0.09, bg:0.24, mean:0.49, Acc.:0.28) improvement than using "weights sharing" strategy.

5) IMPORTANCE OF THE CONTEXT AGGREGATION MODULES

We employ context aggregation modules ($C1^{\uparrow\downarrow\leftarrow\rightarrow}$, $C2^{\searrow\nwarrow\nearrow\swarrow}$) to smooth the label prediction map for inner facial components and the face. By doing this, contexts are explicitly propagated and encoded into feature maps. As shown in Table 8, adopting PrRoI Pooling [36] brings (le:0.94, re:0.95, lips:0.48, tongue:1.09, face:0.68, bg:0.46, mean:0.83, Acc.:0.46) improvement than using [25]. Experimentally, adding $C1^{\uparrow\downarrow\leftarrow\rightarrow}$, and $C2^{\searrow\nwarrow\nearrow\swarrow}$ further improve the score by (le:2.73, re:2.69, lips:3.05, tongue:3.10, mean:2.85, Acc.:0.10) and (face:1.09, bg:1.25, mean:0.21, Acc.:0.82), respectively. Some detailed examples are depicted in Fig.7.

Furthermore, we evaluate three different variants of our context aggregation module ($C1^{\uparrow\downarrow\leftarrow\rightarrow}$,$C2^{\searrow\nwarrow\nearrow\swarrow}$), i.e., ($C2^{\searrow\nwarrow\nearrow\swarrow}$,$C1^{\uparrow\downarrow\leftarrow\rightarrow}$), ($C1^{\uparrow\downarrow\leftarrow\rightarrow}$,$C1^{\uparrow\downarrow\leftarrow\rightarrow}$), and ($C2^{\searrow\nwarrow\nearrow\swarrow}$, $C2^{\searrow\nwarrow\nearrow\swarrow}$) on the TCMID dataset. As the result shows in Table 9, our context aggregation module ($C1^{\uparrow\downarrow\leftarrow\rightarrow}$,$C2^{\searrow\nwarrow\nearrow\swarrow}$) gets the best performance.

TABLE 8. Importance of the context aggregation modules.

Methods	le	re	lips	tongue	face	bg	mean	Acc.
Ours (RoI-align [25])	94.97	94.94	92.98	95.89	97.27	97.81	95.21	97.87
Ours(—,—)	93.18	93.20	90.41	93.88	96.86	97.02	93.51	97.11
Ours(C1,—)	95.64	95.62	93.38	96.90	96.92	97.08	95.69	97.21
Ours(—,C2)	93.56	93.61	91.82	94.05	97.41	97.69	94.09	97.93
Ours(C1,C2)	**95.91**	**95.89**	**93.46**	**96.98**	**97.95**	**98.27**	**96.04**	**98.33**

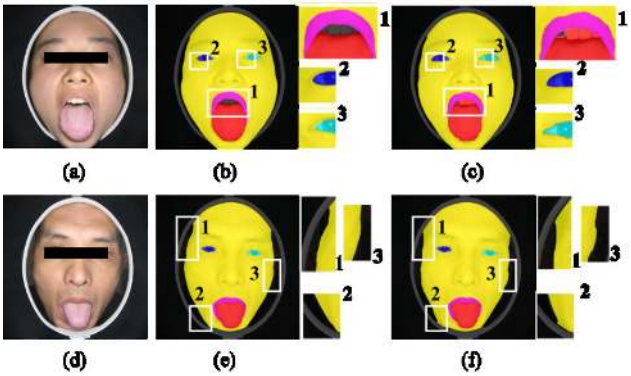

FIGURE 7. Context aggregation modules ($C1^{\uparrow\downarrow\leftarrow\rightarrow}$, $C2^{\searrow\nwarrow\nearrow\swarrow}$) engage useful contexts to local features and model long-range dependencies for segmentation branches Facial Inner Components Segmentation Branch (FICS), Face Segmentation Branch (FS), respectively. (a)(d): input images, (b)(e): output masks of proposed method "Ours(C1,C2)", (c): output mask of "Ours($C1^{\uparrow\downarrow\leftarrow\rightarrow}$,–)", (f): output mask of "Ours(–,$C2^{\searrow\nwarrow\nearrow\swarrow}$)". (best viewed in color) .

TABLE 9. Variant experiments of the context aggregation module.

Methods	le	re	lips	tongue	face	bg	mean	Acc.
Ours(—,—)	93.18	93.20	90.41	93.88	96.86	97.02	93.51	97.11
Ours(C2,C1)	95.73	95.71	93.44	96.88	97.61	98.03	95.87	98.14
Ours(C1,C1)	95.87	95.86	93.44	96.93	97.60	98.04	95.94	98.17
Ours(C2,C2)	95.71	95.72	93.47	96.86	97.90	98.26	95.93	98.21
Ours(C1,C2)	**95.91**	**95.89**	**93.46**	**96.98**	**97.95**	**98.27**	**96.04**	**98.33**

C. COMPARISON WITH STATE-OF-THE-ART METHODS

We perform a thorough comparison between our model and existing state-of-the-art (face parsing and semantic segmentation) methods on LFW-PL⋆, HELEN⋆, and TCMID. In our task, the foreground (le, re, lips, tongue, face) regions are much more important than the background region, so we calculate the mean F-measure over five foreground categories. As Table 10 shows, our model achieves the best results in F-measure over other state-of-the-art (face parsing and semantic segmentation) methods on three datasets (all categories). The average Accuracy, Precision, Recall, F-measure, and their corresponding standard deviations

TABLE 10. The performance results of different methods on various datasets. The performances of each category, together with the mean F-measure over the five foreground categories are listed. Specifically, "mean" indicates the mean F-measure score over the 5 foreground categories (le, re, lips, tongue, face).

Methods	on LFW-PL⋆							on HELEN⋆							on TCMID						
	le	re	lips	tongue	face	bg	mean	le	re	lips	tongue	face	bg	mean	le	re	lips	tongue	face	bg	mean
Smith et al. [14]	77.96	–	69.72	–	80.16	–	75.95	79.17	–	71.83	–	81.77	–	77.59	80.73	–	77.86	–	83.56	–	80.72
Zhou et al. [11]	86.27	–	82.33	–	–	–	84.30	87.79	–	83.56	–	–	–	85.68	87.79	–	84.76	–	–	–	86.28
Wei et al. [31]	83.69	–	–	–	88.12	–	85.91	84.38	–	–	–	89.43	–	86.91	85.19	–	–	–	91.65	–	88.42
Liu et al. [12]	87.08	–	88.38	–	94.37	–	89.94	91.35	–	88.38	–	94.37	–	91.62	91.67	–	88.64	–	95.74	–	92.02
Lin et al. [13]	89.03	89.05	87.97	–	92.74	–	89.70	89.97	90.05	87.91	–	94.68	–	90.65	92.89	92.93	90.63	–	96.37	–	93.21
FCN [24]	74.58	–	63.21	70.17	77.04	79.83	71.25	78.62	–	68.31	73.38	80.27	82.16	75.15	83.12	–	76.51	78.38	84.36	86.92	80.59
CRFasRNN [18]	76.48	–	63.74	71.65	80.96	82.68	73.21	79.35	–	65.70	74.42	83.75	85.11	75.81	83.64	–	76.84	79.62	84.28	87.52	81.10
CNN-CRF [19]	72.65	–	58.37	69.98	81.09	83.60	70.52	74.67	–	60.87	70.52	82.87	85.01	72.23	84.03	–	80.67	83.46	85.03	89.53	83.30
DeeplabV2 [28]	76.31	–	62.11	70.38	83.68	85.31	73.12	78.31	–	63.93	72.08	84.09	85.98	74.60	85.61	–	79.62	82.07	86.36	90.07	83.42
Mask R-CNN [25]	84.48	–	80.41	85.02	–	88.90	83.30	86.09	–	80.69	86.92	–	89.17	84.57	87.56	–	83.69	88.92	–	90.38	86.72
Ours	**92.74**	**92.69**	**91.76**	**94.33**	**95.80**	**96.93**	**93.46**	**93.29**	**93.31**	**91.82**	**95.08**	**96.15**	**96.93**	**93.93**	**95.91**	**95.89**	**93.46**	**96.98**	**97.95**	**98.27**	**96.04**

TABLE 11. The performance results of different methods on three datasets. Average performance metrics (Accuracy, Precision, Recall, F-measure) and the corresponding standard deviations are reported (Average ± Standard Deviations).

Methods	on LFW-PL⋆				on HELEN⋆				on TCMID			
	Accuracy	Recall	Precision	F-measure	Accuracy	Recall	Precision	F-measure	Accuracy	Recall	Precision	F-measure
Smith et al. [14]	78.22±3.19	79.03±3.10	78.03±3.52	78.53±2.86	78.91±2.98	77.96±2.63	78.74±3.02	78.34±2.42	82.73±2.24	81.98±2.33	82.25±2.14	82.11±2.09
Zhou et al. [11]	86.11±2.56	85.64±2.09	86.32±1.88	85.98±2.03	87.14±2.17	86.25±1.67	87.29±2.22	86.77±1.88	87.78±2.13	86.53±2.26	87.29±2.09	86.90±2.02
Wei et al. [31]	86.75±2.67	85.89±2.11	86.03±1.98	85.96±1.86	88.12±1.54	87.24±1.18	87.83±1.67	87.53±1.03	89.13±2.17	88.60±2.32	89.21±1.99	88.90±1.81
Liu et al. [12]	90.21±1.78	89.73±2.13	91.08±1.65	90.40±1.77	92.16±2.03	92.77±2.44	91.89±2.98	92.33±2.51	94.05±1.54	92.04±1.78	94.30±1.26	93.21±1.32
Lin et al. [13]	91.08±1.03	90.34±0.94	91.27±1.15	90.80 ±0.91	93.84±1.25	92.71±1.44	93.58±0.98	93.14±1.03	95.55±1.04	94.20±0.93	95.17±1.02	94.68±0.92
FCN [24]	73.08±4.62	72.74±3.53	72.89±4.02	72.81 ±3.17	77.26±4.02	75.97±3.98	76.48±4.24	76.22±3.88	82.74±3.77	81.21±2.91	81.80 ±3.21	81.50±3.07
CRFasRNN [18]	75.27±3.73	74.23±3.89	75.08±3.21	74.65±3.57	77.94±3.32	75.81±3.24	76.83±3.16	76.32±3.28	83.11±2.98	82.04±2.71	82.86±2.44	82.45±2.31
CNN-CRF [19]	72.16±3.31	71.35±3.77	71.04±3.46	71.19±3.21	74.11±3.77	73.69±3.41	73.98±3.68	73.83 ±3.31	85.23±2.91	83.21±2.77	84.70±2.68	83.95±2.57
DeeplabV2 [28]	75.08±2.11	73.94±2.31	74.26±2.56	74.10±2.10	76.66±3.01	75.49±2.87	76.08±2.79	75.78±2.61	85.72±1.98	82.97±1.69	84.16±1.87	83.56±1.59
Mask R-CNN [25]	85.63±2.09	84.21±1.84	85.12±1.69	84.66±1.35	86.19±1.96	85.71±1.72	86.0±1.84	85.85±1.66	88.05±1.56	86.33±1.32	87.94±1.44	87.13±1.30
Ours	**95.34**±1.01	**94.20**±0.91	**95.07**±1.02	**94.63**±0.87	**95.57**±1.01	**94.69**±0.98	**94.11**±1.22	**94.40**±1.04	**98.33**±0.84	**97.09**±0.92	**97.55**±0.76	**97.32**±0.73

metrics for all the methods on three datasets are displayed in Table 11.

1) COMPARISON WITH FACE PARSING METHODS

As mentioned before, existing face parsing methods [11]–[14], [31] mainly focus on segmenting hair, eyebrows, and other facial components that are rarely relevant to TCMI, rather than segmenting the tongue required for TCMI applications. The F-measure scores of six categories, and the mean F-measures over five foreground categories on three test datasets are presented in Table 10. Our TCMINet achieves the best results over other methods on all categories. As far as the TCMID dataset is concerned, our model achieves the best F-measure of (le:95.91, re:95.89, lips:93.46, tongue:96.98, face:97.95, bg:98.27, mean:96.04), outperforming the state-of-the-art face parsing method [13] by (le:3.02, re:2.96, lips:2.83, tongue:–, face:1.58, bg:–, mean:2.83) score (see Table 10). Table 11 demonstrates further performance comparisons of the proposed method with other existing face parsing approaches. As observed, our method gets the best result on three test datasets in terms of Accuracy, Precision, Recall, and F-measure, which demonstrate the effectiveness of the proposed TCMINet.

2) COMPARISON WITH SEMANTIC SEGMENTATION METHOD

We directly compare the proposed TCMINet with semantic segmentation methods, including FCN [24], CRFasRNN [18], CNN-CRF [19], DeeplabV2 [28], Mask R-CNN [25]. As listed in Table 10, the proposed TCMINet yields a mean F-measure of 96.04, while the mean F-measure of the five competing semantic segmentation methods is 80.59 [24], 81.10 [18], 83.30 [19], 83.42 [28] and 86.72 [25], respectively on TCMID. Futhermore, TCMINet still yields good performance on LFW-PL⋆, HELEN⋆ datasets, which demonstrate the robustness of the proposed TCMINet (see Table10 and Table11). Experimentally, these semantic segmentation methods [18], [19], [24], [25], [28] can't distinguish left and right eyes. As these methods misclassified instances that share similar appearance but have different semantic labels. To be specific, our TCMINet gets larger improvement compared with existing approaches [25], [28] by reducing misclassification errors. In TCMINet, the bounding-box prediction module is more straight-forward but effective for parsing inner facial components. As exemplified in Fig.8, directly applying the region proposal network Mask R-CNN [25] in face parsing causes misclassification problems: the left eyes are recognized as the right eyes.

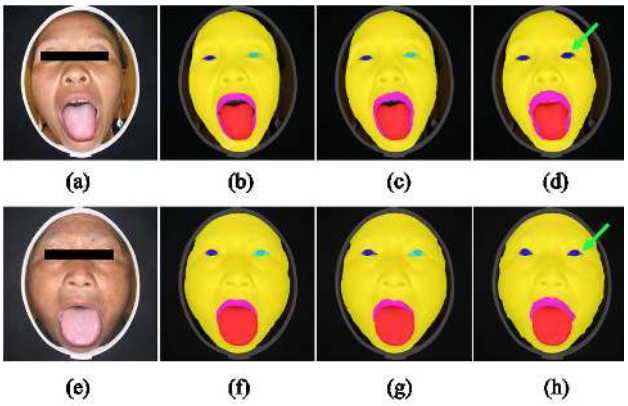

FIGURE 8. Directly applying Mask R-CNN [25] in face parsing causes misclassification problems. (a)(e): input images, (b)(f): groundtruth, (c)(g): output masks of our method, (d)(h): output masks of Mask R-CNN. Misclassification problems are indicated by green arrows. (best viewed in color).

D. QUALITATIVE RESULTS

1) LFW-PL⋆ AND HELEN⋆

We evaluate our approach on LFW-PL⋆ and HELEN⋆ datasets. Experimentally, our model shows a good generalization ability on these two challenging datasets (see Table 10). Fig.9 and Fig.10 show the qualitative parsing results on LFW-PL⋆ and HELEN⋆ dataset, respectively. The ground truth label maps are also shown.

2) TCMID

Our model is robust to challenging inputs. As shown in Fig.11 and Fig.12, the proposed TCMINet is suitable for segmenting face and inner facial components with varying appearances (e.g., tongue substance color, tongue coating color, lip color, facial gloss, and face color) or states (e.g., head rotation, mouth state, eye state, and interference).

VI. DISCUSSION

A. SIMULTANEOUS SEGMENTATION OF MULTIPLE FACE ORGANS

Facial medical analysis is a non-invasive, non-contact diagnostic method of TCM. Generally, segmenting facial skin facial and sensory organs from face images is the first step in computer-aided facial medical analysis. According to related literature, there have been a large number of researches focus on detecting and segmenting single face organ or facial skin. For instance, Pang *et al.* [46] proposed the Bi-Elliptical Deformable Contour (BEDC) model for automated tongue area segmentation. In our previous work [45], we proposed a real-time tongue image segmentation method for remote diagnosis (see Fig13.(b)). Zhao *et al.* [2] develop a facial region segmentation method to partition the facial skin into five specific regions (see Fig13.(c)). In [48], four facial skin blocks in TCM are automatically extracted from each half-face image (see Fig13.(d)). In [47], a cheek region extraction method has been proposed for face diagnosis. (see Fig13.(e)).

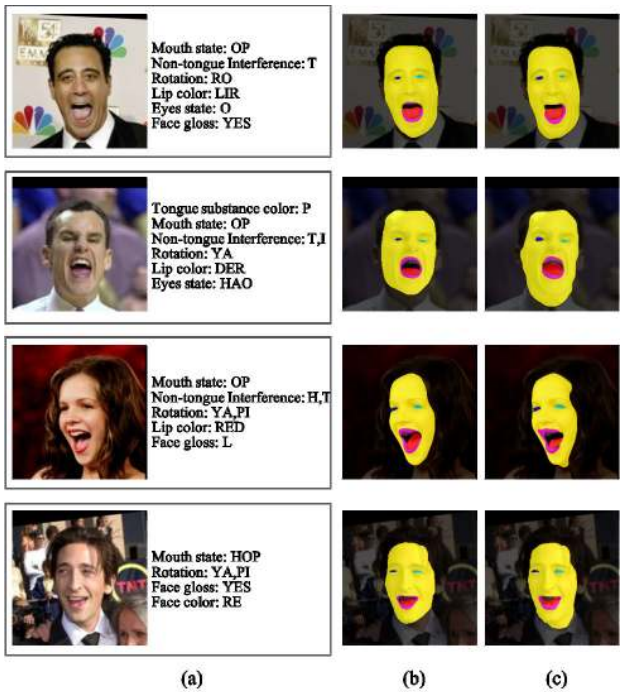

FIGURE 9. Visualizing the results on the LFW-PL⋆ dataset. (a): input images and corresponding attributes, (b): groundtruth, (c): output mask of the TCMINet. (best viewed in color).

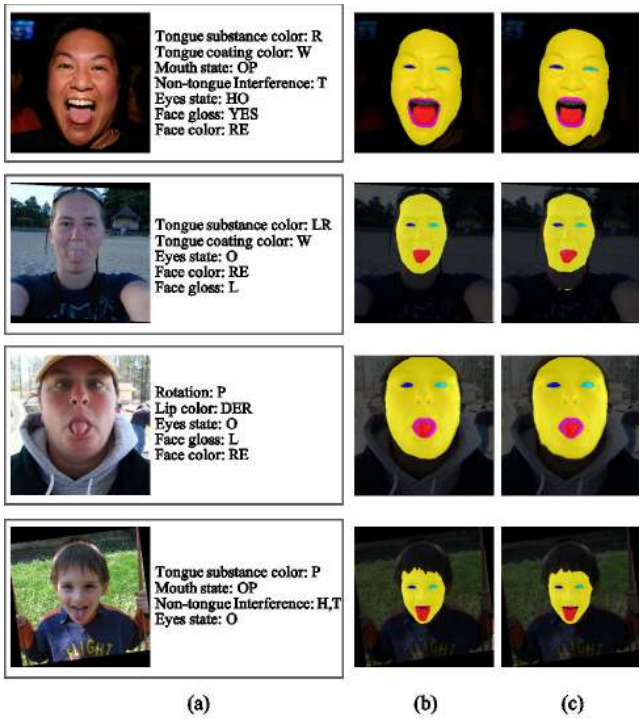

FIGURE 10. Visualizing the results on the HELEN⋆ dataset. (a): input images and corresponding attributes, (b): groundtruth, (c): output mask of the TCMINet. (best viewed in color).

These methods mainly focus on exploring the important role of facial skin regions in reflecting the health status of patients

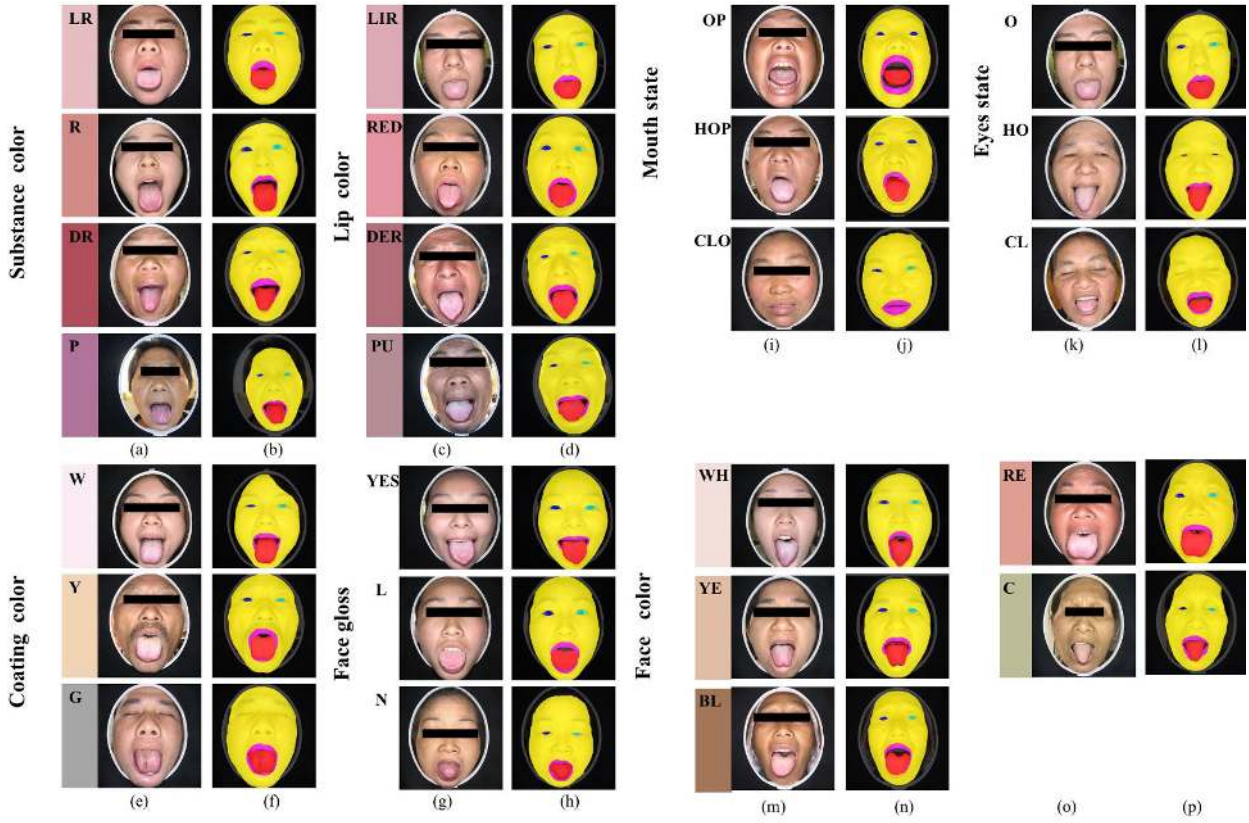

FIGURE 11. Visualization results on challenging images. (a) (c) (e) (g) (i) (k) (m) (o): input images and corresponding attributes, (b) (d) (f) (h) (j) (l) (n) (p) : output mask of the TCMINet. (best viewed in color).

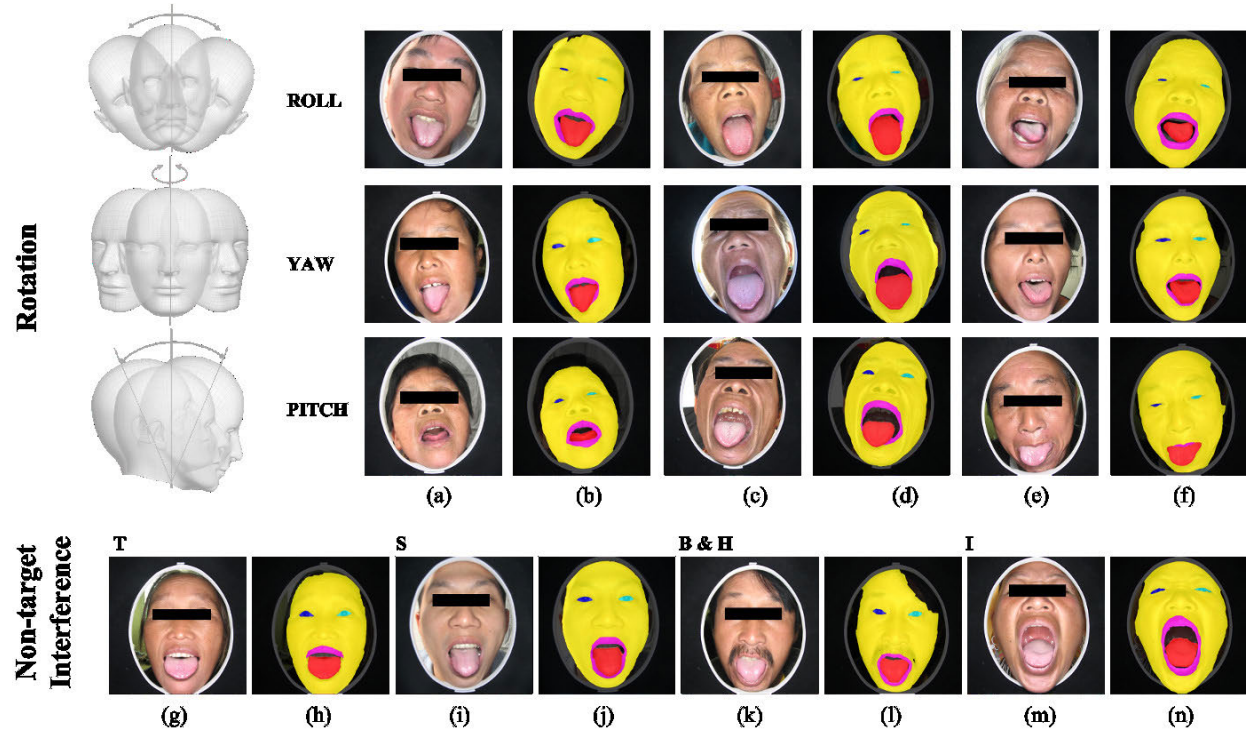

FIGURE 12. Visualization results on challenging images (Head rotation, Interference). (a) (c) (e) (g) (i) (k) (m) : input images, (b) (d) (f) (h) (j) (l) (n) : output mask of the TCMINet. (best viewed in color).

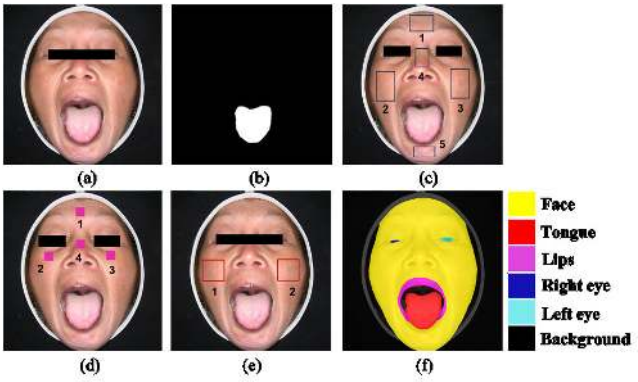

FIGURE 13. Facial medical segmentation methods. (a): Input image. (b): Result of [45]. (c): Result of [2]. Partition of facial skin into five specific regions. (d): Result of [48]. A facial image with four located key blocks. (e): Result of [47]. Two specific regions of facial cheek. (f): output mask of the TCMINet.

while ignoring the criticality of inner facial components (e.g., eyes, lips, tongue). Our work is the first attempt to deal with face and multiple inner facial components simultaneously based on the principles of TCM holistic view (see Fig.13.(f)).

B. FACE PARSING FOR TCMI

As mentioned in SectionI, previous face parsing methods mainly focus on segmenting hair, eyebrows, and other facial components that are rarely relevant to TCMI. In TCMI, the human face and facial sensory components are believed to reveal signs of various constitutions. The tongue among them, as the primary organ of gustation, conveys abundant valuable information about the diseases of the internal body. In our work, face parsing for TCMI amounts to labeling each pixel with the left eye, right eye, lips, tongue, face, and background following the principles of TCM holistic view. Experimentally, our proposed TCMINet outperforms state-of-the-art methods on LFW-PL⋆, HELEN⋆, and TCMID datasets under different evaluation metrics.

C. LIMITATIONS AND FUTURE WORKS

(1) In-the-wild and multi-face conditions: as shown in Table 10 and Table11, our model achieves better performance on the TCMID than on LFW-PL⋆ and HELEN⋆. Although our proposed model is suitable for segmenting faces and inner facial components with varying appearances or states, it cannot deal with multiple faces in field conditions. In future work, we will further extend our architecture to handle different face instances under various environments. (2) Multiple facial specific regions: our proposed model achieves simultaneous segmentation of the face and inner facial components. However, based on the principle of TCM, the human face can be roughly partitioned into multiple specific regions by connecting specific landmarks. Different regions can reflect the health status of different internal organs. In the future, we plan to explore the multi-task learning architecture to achieve multiple facial specific regions partition.

VII. CONCLUSION

In this paper, we propose an effective hybrid network of face parsing for TCMI with context aggregation. Ablation studies show the effectiveness of our hybrid structure and important modules. The superior performances on LFW-PL⋆, HELEN⋆, and the proposed TCMID datasets show the ability of the proposed TCMINet to handle the problem of face parsing for TCMI. Most importantly, our TCMINet can handle faces and all the inner facial components with various appearances, e.g., color, and states, e.g., head rotation, providing new insights into TCMI research and development.

REFERENCES

[1] F. Cheung, "TCM: Made in China," *Nature*, vol. 480, no. 7378, pp. 82–83, Dec. 2011.
[2] C. Zhao, G.-Z. Li, F. Li, Z. Wang, and C. Liu, "Qualitative and quantitative analysis for facial complexion in traditional chinese medicine," *BioMed Res. Int.*, vol. 2014, Dec. 2014, Art. no. 207589.
[3] Y.-M. Li, H.-Z. Yang, W.-B. Guan, Q.-S. Ke, M. Dai, H.-P. Xie, and S.-J. Zhang, "Therapeutic effect of traditional chinese medicine on coagulation disorder and accompanying intractable jaundice in hepatitis b virus-related liver cirrhosis patients," *World J. Gastroenterol.*, vol. 14, no. 39, pp. 6060–6064, 2008.
[4] J. K. Anastasi, M. Chang, J. Quinn, and B. Capili, "Tongue inspection in TCM: Observations in a study sample of patients living with HIV," *Med. Acupuncture*, vol. 26, no. 1, pp. 15–22, Feb. 2014.
[5] F. Li, C. Zhao, Z. Xia, Y. Wang, X. Zhou, and G.-Z. Li, "Computer-assisted lip diagnosis on traditional chinese medicine using multi-class support vector machines," *BMC Complementary Alternative Med.*, vol. 12, no. 1, p. 127, Dec. 2012.
[6] Y. Hu, H. Lu, J. Cheng, W. Zhang, F. Li, and W. Zhang, "Robust lip segmentation based on complexion mixture model," in *Advances in Multimedia Information Processing*, E. Chen, Y. Gong, and Y. Tie, Eds. Cham, Switzerland: Springer, 2016, pp. 85–94.
[7] X. Li, T. Yang, Y. Hu, M. Xu, W. Zhang, and F. Li, "Automatic tongue image matting for remote medical diagnosis," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Nov. 2017, pp. 561–564.
[8] P. Rot, Z. Emersic, V. Struc, and P. Peer, "Deep multi-class eye segmentation for ocular biometrics," in *Proc. IEEE Int. Work Conf. Bioinspired Intell. (IWOBI)*, Jul. 2018, pp. 1–8.
[9] J. Qiu, "Traditional medicine: A culture in the balance," *Nature*, vol. 448, no. 7150, pp. 126–128, 2007.
[10] R. Yuan, W.-L. Shi, Q.-Q. Xin, K.-J. Chen, and W.-H. Cong, "Holistic regulation of angiogenesis with chinese herbal medicines as a new option for coronary artery disease," *Evidence-Based Complementary Alternative Med.*, vol. 2018, pp. 1–10, Aug. 2018.
[11] Y. Zhou, X. Hu, and B. Zhang, "Interlinked convolutional neural networks for face parsing," in *Proc. Int. Symp. Neural Netw.*, 2015, pp. 222–231.
[12] S. Liu, J. Shi, L. Ji, and M.-H. Yang, "Face parsing via recurrent propagation," in *Proc. Brit. Mach. Vis. Conf.*, 2017, pp. 1–13.
[13] J. Lin, H. Yang, D. Chen, M. Zeng, F. Wen, and L. Yuan, "Face parsing with RoI tanh-warping," 2019, *arXiv:1906.01342*. [Online]. Available: http://arxiv.org/abs/1906.01342
[14] B. M. Smith, L. Zhang, J. Brandt, Z. Lin, and J. Yang, "Exemplar-based face parsing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3484–3491.
[15] P. Luo, X. Wang, and X. Tang, "Hierarchical face parsing via deep learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2480–2487.
[16] B. Shuai, Z. Zuo, B. Wang, and G. Wang, "Scene segmentation with DAG-recurrent neural networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1480–1493, Jun. 2018.
[17] B. Shuai, Z. Zuo, B. Wang, and G. Wang, "DAG-recurrent neural networks for scene labeling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3620–3629.

[18] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. S. Torr, "Conditional random fields as recurrent neural networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1529–1537.

[19] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," *Comput. Sci.*, vol. 4, pp. 357–361, Oct. 2014.

[20] A. Kae, K. Sohn, H. Lee, and E. Learned-Miller, "Augmenting CRFs with Boltzmann machine shape priors for image labeling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2019–2026.

[21] F. Visin, K. Kastner, K. Cho, M. Matteucci, A. Courville, and Y. Bengio, "ReNet: A recurrent neural network based alternative to convolutional networks," 2015, *arXiv:1505.00393*. [Online]. Available: http://arxiv.org/abs/1505.00393

[22] F. Visin, A. Romero, K. Cho, M. Matteucci, M. Ciccone, K. Kastner, Y. Bengio, and A. Courville, "ReSeg: A recurrent neural network-based model for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2016, pp. 426–433.

[23] X. Ding, Y. Jiang, X. Qin, Y. Chen, W. Zhang, and L. Qi, "Reading face, reading health: Exploring face reading technologies for everyday health," in *Proc. Conf. Hum. Factors Comput. Syst.*, 2019, p. 205.

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

[25] K. He, G. Gkioxari, P. Dollar, and R. B. Girshick, "Mask R-CNN," in *Proc. Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.

[26] 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.

[27] Z. Huang, L. Huang, Y. Gong, C. Huang, and X. Wang, "Mask scoring R-CNN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6409–6418.

[28] 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," 2016, *arXiv:1606.00915*. [Online]. Available: http://arxiv.org/abs/1606.00915

[29] X. Ou, S. Liu, X. Cao, and H. Ling, "Beauty eMakeup: A deep makeup transfer system," in *Proc. ACM Multimedia Conf.*, 2016, pp. 701–702.

[30] H. Zhang, B. S. Riggan, S. Hu, N. J. Short, and V. M. Patel, "Synthesis of high-quality visible faces from polarimetric thermal faces using generative adversarial networks," *Int. J. Comput. Vis.*, vol. 127, nos. 6–7, pp. 845–862, Jun. 2019.

[31] Z. Wei, Y. Sun, J. Wang, H. Lai, and S. Liu, "Learning adaptive receptive fields for deep image parsing network," *Comput. Vis. Pattern Recognit.*, vol. 4, no. 3, pp. 231–244, 2017.

[32] L. Zhou, Z. Liu, and X. He, "Face parsing via a fully-convolutional continuous CRF neural network," 2017, *arXiv:1708.03736*. [Online]. Available: http://arxiv.org/abs/1708.03736

[33] N. Kalchbrenner, I. Danihelka, and A. Graves, "Grid long short-term memory," *Neural Evol. Comput.*, 2015. [Online]. Available: https://arxiv.org/abs/1507.01526

[34] A. Graves, S. Fernandez, and J. Schmidhuber, "Multi-dimensional recurrent neural networks," *Proc. Int. Conf. Artif. Neural Netw.*, 2007, pp. 549–558.

[35] A. van den Oord, N. Kalchbrenner, and K. Kavukcuoglu, "Pixel recurrent neural networks," *Comput. Vis. Pattern Recognit.*, 2016. [Online]. Available: https://arxiv.org/abs/1601.06759

[36] B. Jiang, R. Luo, J. Mao, T. Xiao, and Y. Jiang, "Acquisition of localization confidence for accurate object detection," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 816–832.

[37] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1800–1807.

[38] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 764–773.

[39] L. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 833–851.

[40] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1026–1034.

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

[42] X. Shen, X. Tao, H. Gao, C. Zhou, and J. Jia, "Deep automatic portrait matting," in *Computer Vision*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham, Switzerland: Springer, 2016, pp. 92–107.

[43] A. Quattoni and A. Torralba, "Recognizing indoor scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 413–420.

[44] J. Herve, D. Matthijs, and C. Schmid, "Hamming embedding and weak geometry consistency for large scale image search," *Proc. 10th Eur. Conf. Comput. Vis.*, 2008, pp. 413–420.

[45] X. Li, D. Yang, Y. Wang, S. Yang, L. Qi, F. Li, Z. Gan, and W. Zhang, "Automatic tongue image segmentation for real-time remote diagnosis," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Nov. 2019, pp. 409–411.

[46] 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.

[47] Y. Yang, J. Zhang, L. Zhuo, Y. Cai, and X. Zhang, "Cheek region extraction method for face diagnosis of traditional chinese medicine," in *Proc. IEEE 11th Int. Conf. Signal Process.*, Oct. 2012, pp. 1663–1667.

[48] T. Shu, B. Zhang, and Y. Yan Tang, "An extensive analysis of various texture feature extractors to detect diabetes mellitus using facial specific regions," *Comput. Biol. Med.*, vol. 83, pp. 69–83, Apr. 2017.

XINLEI LI is currently pursuing the Ph.D. degree with the Academy for Engineering and Technology, Fudan University, Shanghai, China. Her current research interests include semantic segmentation, medical image analysis, machine learning, and deep learning.

DAWEI YANG is currently pursuing the Ph.D. degree with the Academy for Engineering and Technology, Fudan University, Shanghai, China. His current research interests include computer vision, AI chip architecture, and heterogeneous programming.

YAN WANG is currently pursuing the Ph.D. degree with the Academy for Engineering and Technology, Fudan University, Shanghai, China. His current research interests include computer vision, image processing, quality of experience, machine learning, and deep learning.

WEI ZHANG received the B.A. and M.A. degrees in economics and the Ph.D. degree in computer science from Fudan University, Shanghai, China, in 2000, 2003, and 2008, respectively. He is currently an Associate Professor with the School of Computer Science, Fudan University. He is also a Visiting Scholar with the Computer Science Department, University of North Carolina at Charlotte, from 2016 to 2017. His current research interests include statistical pattern recognition, deep neural networks, and video object segmentation.

FUFENG LI is currently a Professor with the Shanghai University of Traditional Chinese Medicine. She has been engaged in the standardization and objectification of the four diagnostic methods of TCM. She is also a Registered Expert of the ISO/TC249 International Standard of TCM, the Standing Director of the Professional Committee of TCM Diagnosis Instruments with the World Federation of Traditional Chinese Medicine, and a member of the Professional Committee of TCM Diagnosis with the Chinese Society of TCM.

WENQIANG ZHANG was born in Zhaoyuan, Shandong, China, in 1970. He received the B.S. degree in electric power engineering from the Huazhong University of Science and Technology, Wuhan, China, in 1992, the M.S. degree from Shandong University, Jinan, China, in 2001, and the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China, in 2004, all in mechanical engineering. From 2004 to 2007, he was an Assistant Professor with the Computer Engineering Department. From 2008 to 2016, he was an Associate Professor with the School of Computer Science, Fudan University, where he has been a Professor since 2017.