# Fine-Grained Lip Image Segmentation using Fuzzy Logic and Graph Reasoning

Lei Yang, Shilin Wang*, *Senior Member, IEEE*, and Alan Wee-Chung Liew, *Senior Member, IEEE*

***Abstract*— Fine-grained lip image segmentation plays a critical role in downstream tasks such as automatic lipreading, as it enables the accurate identification of inner mouth components such as teeth and tongue which are essential for comprehending spoken utterances. However, achieving accurate and robust lip image segmentation in natural scenes is still challenging due to significant variations in lighting condition, head pose and background. This paper proposes a novel deep neural network based method for fine-grained lip image segmentation that exploits fuzzy and graph theories to handle these variations. A fuzzy learning module is designed to deal with the uncertainties in color and edge information and enhance feature maps at various scales. The fuzzy graph reasoning module with fuzzy projection models the relationship among semantics components and achieves a global receptive field. In our experiments, a fine-grained lip region segmentation dataset, i.e., FLRSeg, is built for evaluation and experiment results have shown that the proposed method can achieve superior segmentation performance (94.36% in pixel accuracy and 74.89% in mIoU) compared with several SOTA lip image segmentation methods.**

***Index Terms*—fuzzy neural networks, convolutional neural network, lip image segmentation, graph reasoning**

## I. Introduction

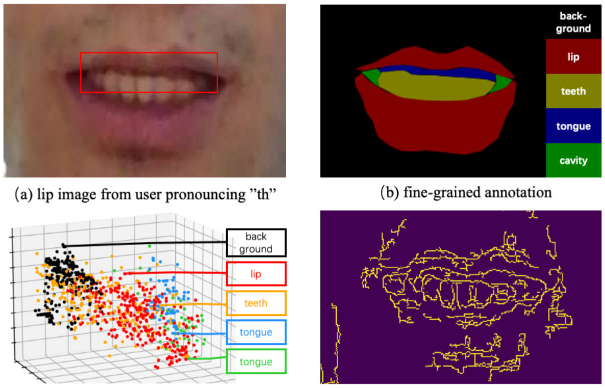
Lip image segmentation, also referred to as lip segmentation, aims to provide the content label at a pixel level for an image containing the lip region. The content labels usually contain the lip and background, and the lip image segmentation results can be used in many downstream applications, including automatic lipreading [1], [2], visual speaker identification and authentication [3]–[5], lip synchronization for facial animation [6], etc. Hence, it has attracted widespread research interests in the past decades.

In early years, the traditional lip image segmentation approaches can be roughly divided into three categories: color-based, edge-based, and spatial information guided methods. Color-based methods [7], [8] segmented the lip region by a preset color filter. Edge-based approaches [9], [10] employed edge/gradient information to extract the lip contour. To guarantee the extracted lip contour can form a valid lip shape, ASM [11] and AAM models [12] were the two widely used lip models in edge-based approaches. The spatial information guided methods [13]–[17] assume that the lip pixels are usually clustered together to form a large patch. Both the local spatial information [13], [14], i.e., the neighboring consistency, and the global spatial information [15]–[17] can be incorporated to improve the robustness of the segmentation results. These methods can achieve reliable segmentation results in the laboratory scenario where the lighting condition and background do not change much.

With the rapid development of deep learning theory, deep neural networks (DNN) have achieved outstanding performance in many computer vision tasks. The fully convolutional network (FCN) [18] based methods with the encoder-decoder structure have been successfully applied in lip image segmentation. The encoder, which usually consists of several stacked convolutional layers, extracts multi-scale features in a feature pyramid [19]. To increase the receptive field, some methods introduced dilated [20] or deformable [21] convolutions in the encoder part, and other methods [22], [23] used spatial pyramid pooling to capture multi-scale feature maps in the decoder part. The attention mechanism [24], [25] were also used to expand the receptive fields. In [26], a Lip Segmentation Network (LSN) was designed to classify lip pixels in images, which adopted the classical FCN structure and integrated additional information from neighboring frames in a lip image sequence. In a recent work by our group [27], the lip segmentation with a fuzzy convolutional neural network (LSFCNN) was proposed. A fuzzy learning module was designed and seamlessly integrated into the deep neural network to handle color/edge uncertainties in the open mouth scenario.

Using prior information in hand-labeled training samples, the sophisticated DNN-based lip image segmentation methods above have outperformed traditional methods in both accuracy and robustness. It is worth noting that most existing lip image segmentation methods focus on differentiating lip pixels from the background, and the inner mouth components such as teeth or tongue are usually regarded as background. However, these components are highly related to the pronunciation process [28] and can enhance the performance of downstream applications like lipreading. Therefore, a fine-grained lip image segmentation is needed to provide pixel-level annotations for both lip and inner mouth components. Nevertheless, accurate and robust fine-grained lip image segmentation is very challenging due to: i) great variations caused by illumination, head pose, etc. in natural scenes and ii) high similarity in color/edge/spatial information between the lip pixels and pixels belonging to some inner mouth components such as tongue or gums. To address these challenges, a new deep

Lei Yang and Shilin Wang are with the School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, 200240, Shanghai, China. (e-mail: yangleisx@sjtu.edu.cn, wsl@sjtu.edu.cn). A.W.C. Liew is with the School of ICT, Griffith University, QLD 4222, Australia (email: a.liew@griffith.edu.au). * Shilin Wang is the corresponding author. The work described in this paper was supported in part by the National Natural Science Foundation of China (62271307, 61771310).



(a) lip image from user pronouncing "th"

(b) fine-grained annotation

(c) color distribution in LAB color space

(d) edge map detected by Canny operator

Fig. 1. (a) The original lip image where the tongue pixels in the red rectangle are difficult to differentiate; (b) Human annotations; (c) Color distribution of various components; (d) Edge maps obtained by the Canny operator.

neural network structure is proposed for fine-grained lip image segmentation. The proposed model incorporates a fuzzy learning module and a fuzzy graph reasoning module are proposed to handle uncertainties in color, edge, and annotations, thereby improving the accuracy and robustness of the segmentation network. The main contributions of our approach are three-folds:

1) A novel architecture is proposed to incorporate fuzzy logic and graph reasoning into DNN for fine-grained lip image segmentation, which can effectively exploit the color, edge, spatial information and address the challenges posed by variations resulting from illumination, image noise, and hand-labeled annotations. To the best of our knowledge, there are very few works that address the challenging task of inner mouth component segmentation under natural scenes.
2) A new fuzzy learning module is proposed that can be seamlessly integrated into neural networks for complex fuzzy rules modeling. Furthermore, a fuzzy graph reasoning module with inline fuzzy clustering is proposed to expand the receptive field, capture global information, and model relations between objects.
3) Experiments on a fine-grained lip image segmentation dataset have demonstrated that the proposed approach outperforms SOTA methods with slight overhead.

## II. Chanllenges and Motivations

### A. Problem Description in Fine-Grained Lip Image Segmentation

Fine-Grained lip image segmentation in natural scenes is a challenging task especially when the mouth is open and several inner mouth components (e.g., teeth, tongue, oral cavity, etc.) are visible [17]. Fig. 1 shows an example. The following observations can be made from the figure: i) Compared with the traditional lip image segmentation problem with two classes (lip vs. background), fine-grained lip image segmentation with multiple object classes (lip, teeth, tongue, etc.) encounters much more serious color overlapping problems between different classes (e.g. lip vs. tongue, teeth vs. lip, etc. as shown in Fig. 1c); ii) As shown in Fig. 1d, the edge map is quite complex and there is no obvious boundaries between different classes; iii) Since the spatial appearance of inner mouth components (e.g. tongue) varies greatly, the spatial information contains little class-related information; iv) Considering large variations in color, edge and spatial information, pixel-level segmentation using simple, handcrafted rules cannot provide reliable results. For DNN-based methods that can extract complex rules from the above information, accurate and consistent pixel-level annotation becomes a new challenge. As shown in Fig. 1a, different annotator may get different result in the inner mouth region, which will confuse the feature extraction and classification layers in DNN. Due to the above issues, multi-class pixel-level lip image segmentation under natural scenes is still an open problem.

### B. Lip Segmentation by DNN with Fuzzy Logic and Graph Reasoning

Previous works [26] have demonstrated that compared with the handcrafted features, CNN features can better depict the relationship between the color/spatial information of a pixel and its corresponding lip/background class label. However, in the multiple class classification task of fine-grained lip image segmentation, due to the great variations/uncertainties in color, edge, spatial location, training sample annotation, etc., CNN features alone cannot achieve reliable results. In order to improve the accuracy and robustness of the segmentation method, a fuzzy learning module and a fuzzy graph reasoning module are designed and seamlessly incorporated into the newly designed deep neural network.

Fuzzy systems have been proven to be effective in handling data that is uncertainties and ambiguities [29]–[32]. In our previous work [27], we demonstrated that a fuzzy learning module using the "AND" fuzzy logic can improve the discriminative power of the DNN features in the traditional lip vs background classification problem. In face of the new challenges in fine-grained lip image segmentation, we extend the idea in [27] and design a new fuzzy learning module. This new module employs fuzzy rules to extract features with high discriminative power while reduce the influence of unrelated features. With the fuzzy rules, the new fuzzy learning module can learn the complex rules around inner mouth components. Besides, the output feature maps are reconstructed from the fuzzy rules and focus on the relevant features. Compared to our previous work which directly use the firing strength as output of the fuzzy module, the proposed fuzzy learning module aggregates features that are sampled from the learned Gaussian kernel in fuzzifier stage according to the firing strength of each fuzzy rule. This approach not only enhances the feature maps but also ensures that the entire module is differentiable and friendly for end-to-end learning.

On the other hand, most existing lip image segmentation networks adopt FCN with a hierarchical feature extraction structure. The extracted features for each pixel are used to represent the color information of the corresponding pixel and its relationship among neighboring pixels (determined

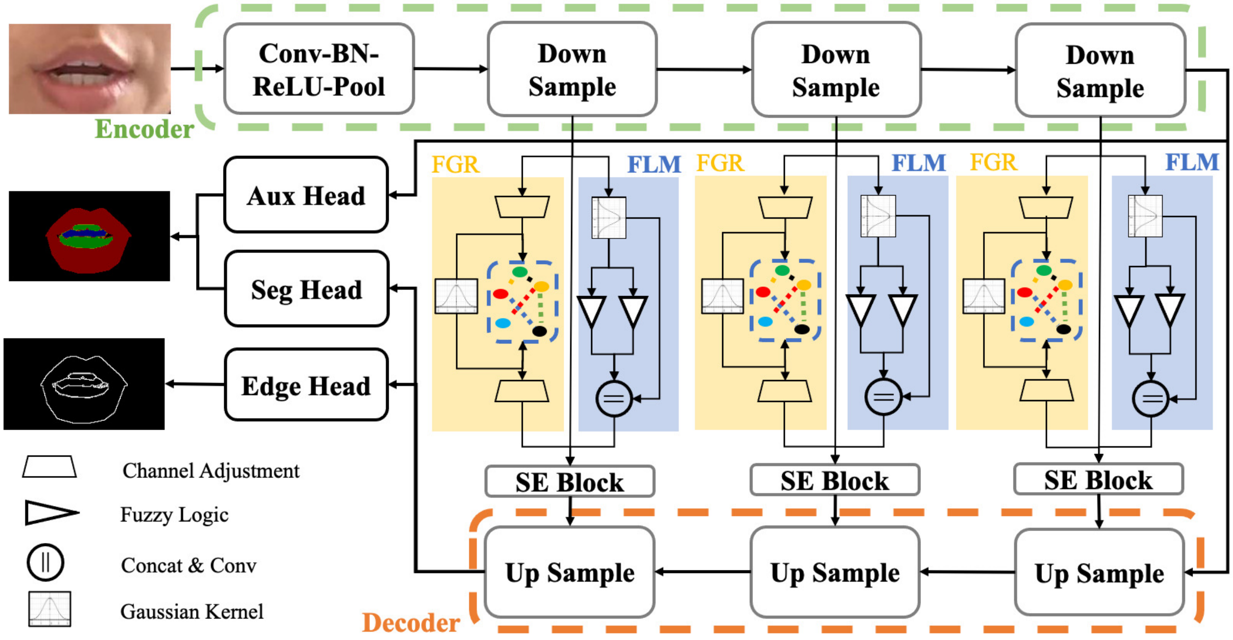


Fig. 2. The overall architecture of the proposed network.

by the kernel size). However, the great variations in fine-grained lip image segmentation can degrade the discriminative power and robustness of these features. Additional information is therefore required to refine the feature map. Intuitively, there are intrinsic relationships among the features for various sematic classes. For example, none of the inner mouth components is visible for a closed mouth image, and gum pixels are located nearby the teeth and lip. Considering these relationships among sematic components (lip, teeth, and tongue) can provide a semantic view of the entire lip image at the pixel level.

To incorporate this sematic information, we propose a fuzzy graph reasoning module. In this module, pixels in the image are transformed into abstract nodes to construct graphs in features space and a simplified Graph Convolutional Network (GCN) is adopted to reason over the graph features. During the construction of the graph, features for pixels from the whole image are taken into account, including both local information such as color gradient and location, and global information. At the same time, a fuzzy clustering layer is adopted as the transformation matrix which build up fuzzy relations from pixel features to graph nodes, which is able to deal with the uncertainties in the image.

The fuzzy learning module and the fuzzy graph reasoning modules are the two key components that enable accurate and robust segmentation results for fine-grained lip image segmentation. They can refine the final feature map from two different aspects and effectively address the major problems discussed in subsection II-A.

## III. Method

### A. Overall Structure

The overall network structure of the proposed network follows the U-Net [33] schema, as shown in Fig. 2. The model is composed of an encoder, a feature enhancement module, and a decoder, followed by three prediction heads. The feature enhancement module consists of the fuzzy learning module (FLM) and the fuzzy graph reasoning module (FGR). Details are presented in the following subsections.

In the encoder, a deep convolution neural network is adopted to extract multi-scale features. In the decoder, feature maps and outputs from the fuzzy learning modules and fuzzy graph reasoning modules are fused together with a Squeeze-and-Excitation (SE) block [24] at each scale. Then, the fused feature maps are fed to the segmentation head to generate the final segmentation result. Note that an edge detection head is used in parallel to the final segmentation head in a multi-task learning schema to improve the segmentation performance.

### B. Fuzzy Learning Module

The objective of the fuzzy learning module is to model the complex rules between the feature map and the semantics categories of every pixel. For each feature map at a particular scale, a fuzzy learning module is adopted, whose structure is shown in Fig. 3. The proposed fuzzy learning module can be divided in to three parts.

*1) Fuzzifier:* Firstly, a group of fuzzy functions is applied to transform feature maps extracted by the encoder into fuzzy values, which is a float number that indicates how the features match the fuzzy kernels.

Let $F$ be the feature map extracted in the encoder with a size of $H \times W \times C$, where $H$, $W$ and $C$ denote the height, width, and number of channels of the feature map, respectively. As shown in Fig. 3, for each channel, a group

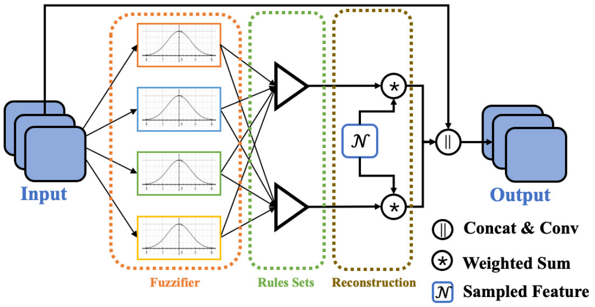


Fig. 3. The overall structure of the proposed Fuzzy Learning Module (FLM).

of membership functions is applied to calculate the fuzzy membership values over $K$ different kernels. Each kernel function is in the Gaussian form formulated in (1) and it assigns a fuzzy linguistic term label to the feature point according to the feature map and the fuzzy kernel.

$$M_{x,y,c,k} = e^{-\left(\frac{X_{x,y,c}-\mu_{c,k}}{\sigma_{c,k}}\right)^2} \tag{1}$$

where $x = 1 \ldots H, y = 1 \ldots W$ is the spatial coordinate of the feature point $X_{x,y,c}$ in channel $c$, $\mu_{c,k}$ and $\sigma_{c,k}$ are the mean and standard deviation of the k-th membership function, $k = 1 \ldots K$. $M_{x,y,c,k}$ denotes the k-th membership value of the feature point $X_{x,y,c}$. Note that $\mu_{c,k}$ and $\sigma_{c,k}$ are randomly initialized and learnable. During training over the dataset, the membership functions are tuned to capture significant linguistic terms from the feature maps.

By fuzzifying the feature maps, the model is capable of handling the noise and uncertainties in images using non-exclusive fuzzy membership values, which are more robust than deterministic features.

*2) Fuzzy Logic:* After fuzzifying the feature maps, two sets of fuzzy rules are calculated over the linguistic terms for feature reconstruction. In order to extract features with high discriminative power and reduce the influence of unrelated features, the fuzzy logic focuses on the most and least well-matched semantics categories.

One set of fuzzy rules are formulated in (2), in which the most well-matched linguistic term is selected. The output weight value will be greater only when the membership value is greater.

$$\begin{aligned} R_i^1 :& IF\ x_i\ is\ K_1\ OR\ x_i\ is\ K_2\ OR\ \ldots \\ & THEN\ w_i^1 = \arg\max M_k \end{aligned} \tag{2}$$

where $x_i$ is a point of the input feature, $i = 1, 2, \ldots, H \times W \times C$. In the proposed fuzzy learning module, the OR logic is calculated through MAX operation, and the firing strength of the first set of fuzzy rules is the maximum values over the feature point.

The other set of fuzzy rules are formulated in (3), which is complementary to the first set of rules. In the second set of rules, the least well-matched linguistic term is selected where the output weight value is greater only when the membership value is smaller.

$$\begin{aligned} R_i^2 :& IF\ x_i\ is\ not\ K_1\ OR\ x_i\ is\ not\ K_2\ OR\ \ldots \\ & THEN\ w_i^2 = \arg\max M_k \end{aligned} \tag{3}$$

To simplify calculation and ensure that all the parameters can be optimized using end-to-end training with gradient descent, the fuzzy logic is implemented as a scaled softmax operation as in (4) and (5).

$$w^1_{x,y,c,k} = \frac{\exp(-M_{x,y,c,k}/T)}{\sum_k \exp(-M_{x,y,c,k}/T)} \tag{4}$$

$$w^2_{x,y,c,k} = \frac{\exp(M_{x,y,c,k}/T)}{\sum_k \exp(M_{x,y,c,k}/T)} \tag{5}$$

The hyperparameter $T$ in the scaled softmax operations controls the smoothness of the output. In order to make the behavior of the proposed fuzzy logic operation approximate the original logic calculated with the "MIN" and "MAX" operation, the $T$ value is empirically set to 0.05.

*3) Reconstruction:* In this stage, the output feature maps are reconstructed by a weighted summation over features that are sampled from the distribution determined by the parameter in the fuzzifier. By reconstructing the feature map using the output of the fuzzy logic, the model can focus on the most related features and reduce the interference of noise.

During the generation of the enhanced feature map, a set of random variables $F_{c,k}$ are sampled from the Gaussian distribution determined by the membership functions, which is given in (6).

$$F_{c,k} \sim \mathcal{N}(\mu_{c,k}, \sigma^2_{c,k}) \tag{6}$$

The enhanced feature maps are generated by the weighted summation over the sampled features maps, as given in (7) and (7), where the weights are the outputs of the fuzzy logic stage in (4) and (5).

$$Y^1_{x,y,c} = \sum_k w^1_{x,y,c,k} \times F_{c,k} \tag{7}$$

$$Y^2_{x,y,c} = \sum_k w^2_{x,y,c,k} \times F_{c,k} \tag{8}$$

The fuzzy learning module adopts fuzzy logic and feature reconstruction to capture the most salient and least significant components in the feature map based on linguistic terms. The fuzzy rules are implemented using the softmax operation, which guarantees differentiability and also performs weight normalization for the subsequence feature reconstruction part.

Finally, the enhanced feature maps and the original feature maps are concatenated and fed into a convolution layer as shown in (9).

$$Y = Conv(Concat[X, Y^1, Y^2]) \tag{9}$$

The number of learnable parameters in the fuzzy learning module is $C \times K \times 2$, which is comparable to that in a general convolution layer. The number of membership functions K is the only hyperparameter in the module. Note that bottleneck layers with kernel size $1 \times 1$ are adopted for dimension adjustment. One convolution layer is used to compress channel numbers to a quarter of the input channels to reduce the computation cost. Another layer is incorporated in the end of the module so that the output feature map has the same dimension as the input feature map. In this way, the module can be easily plugged into any existing model to integrate fuzzy logic information.

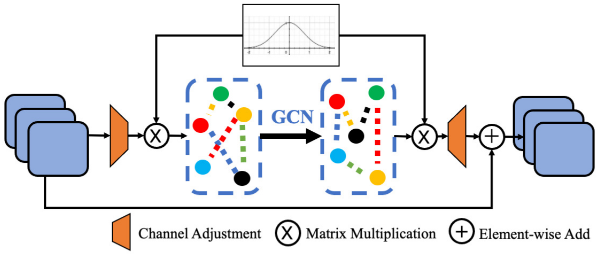


Fig. 4. The overall structure of the proposed Fuzzy Graph Reasoning (FGR) module.

### C. Fuzzy Graph Reasoning Module

The objective of fuzzy graph reasoning is to extract global information from the entire image. For each feature map at a particular scale, a fuzzy graph reasoning module is employed. To address the intense noise and uncertainties in fine-grained lip segmentation, a fuzzy clustering layer is used to build a projection matrix that provides a robust transformation between feature maps and graph nodes. The proposed fuzzy graph reasoning (FGR) module consists of two parts: fuzzy projection and reasoning through GCN, as shown in Fig. 4.

*1) Fuzzy Projection:* In the fuzzy graph reasoning module, a fuzzy projection is used to transform feature maps from pixel space onto the inter-space where each node represents a semantic object in the input image.

To handle uncertainties in the image, fuzzy membership functions are used to extract the semantic relationship between features and graph nodes in the inter-space. The fuzzy functions calculate membership values according to the L2-distance between the pixels in the feature map and the nodes in the graph in a Gaussian form formulated in (10). Each fuzzy function represents a node in the inter-space, and the membership values indicate how much the pixels belongs to the node in a non-exclusive way.

$$M_{d,n} = e^{-\frac{||X_d - \mu_n||^2}{\sigma_n^2}} \tag{10}$$

where the inputs are reshaped as $X \in \mathbb{R}^{D \times C}$, the total number of feature points is $D = H \times W$ and the membership matrix is $M \in \mathbb{R}^{D \times N}$, with number of nodes in graph is $N$.

The projection and re-projection operation are conducted through matrix multiplication, which can be regarded as an inline fuzzy clustering operation. The node in the graph can be seen as centroids of the clusters and the node features are constructed through linear combination. During the transformation, all pixels in the image are considered.

*2) Reasoning by Simplified GCN:* The general graph convolution is given by (11), which is computationally expensive with matrix multiplications.

$$H^{(l+1)} = \sigma(AH^{(l)}W) \tag{11}$$

where $H^{(l)} \in \mathbb{R}^{N \times C}$ is the feature map in l-th layer, $A \in \mathbb{R}^{N \times N}$ is the adjacency matrix of the nodes in graph, $W \in \mathbb{R}^{C \times C}$ is the convolutional weights, and $\sigma(\cdot)$ is the activation function, which is usually the Sigmoid function in GCN.

The general graph convolution operation can be divided into two stages [34]: information exchange between nodes and information exchange between features of each node. In the first stage, each node gathers information from nearby nodes through the adjacent matrix. In the second stage, each node updates itself with a linear transformation. The entire process can be implemented by applying two linear transformations along different dimension, i.e., node-wise and channel-wise, as shown in (12). With linear transformations, the graph convolution is conducted over a fully connected graph.

$$H^{(l+1)} = Linear(Linear(H^{(l)})^T)^T \tag{12}$$

As multi-layer GCNs may lead to over-smoothing [35], the proposed module adopts a two-layer GCN with shared parameters and the same sigmoid activation.

The number of parameters in the fuzzy graph reasoning module is $D \times N \times 2$, and the number of nodes in the graph is the only hyperparameter. Bottleneck layers are applied in the module to reduce the number of channels, and the output of GCN block are projected back to the pixel-space which is in the same shape as the input features after channel extension.

### D. Loss Function Design

The loss function during training is consisted of three parts and each part is employed to supervise three kinds of prediction heads, i.e., the segmentation head, the edge detection head, and the auxiliary head, as shown in Fig. 2. The total loss is calculated as in (13), where $k_{edge}$ and $k_{aux}$ are weights for loss balance.

$$L = L_{seg} + k_{aux} \cdot L_{aux} + k_{edge} \cdot L_{edge} \tag{13}$$

For the segmentation head, a hybrid loss function of cross-entropy and dice loss is adopted to supervise the segmentation result as in (14).

$$L_{seg} = L_{ce} + k_{dice} \cdot L_{dice} \tag{14}$$

The pixel-wise cross-entropy loss formulated in (15) is adopted to supervise the predicted probability of pixels belong to semantics categories, which is widely used in classification tasks

$$L_{ce} = -\sum_{i}^{hw} \sum_{n}^{N} y_{i,n} \log y'_{i,n} \tag{15}$$

where $y'_{i,n}$ means the predicted probability of the i-th pixels classified into the n-th class and $y_{i,n}$ is the ground truth. The dice loss [36], based on dice coefficient, formulated in (16), can deal with the imbalance between foreground and background pixels. Therefore, the dice loss is commonly adopted in semantics segmentation. The combination of cross entropy loss and dice loss is helpful in training a sementation model.

$$L_{dice} = \sum_{n}^{N} (1 - \frac{2 \times |m'_n \cap m_n|}{|m'_n| + |m_n|}) \tag{16}$$

where $m'_n$is predicted mask of the n-th class and $m_n$ is the corresponding ground truth.

For the edge detection head, a binary cross-entropy loss $L_{edge}$ in (17) is adopted to supervise the edge detection result, which has been proven [37], [38] that can improve segmentation through multi-task learning. The ground truth for edge

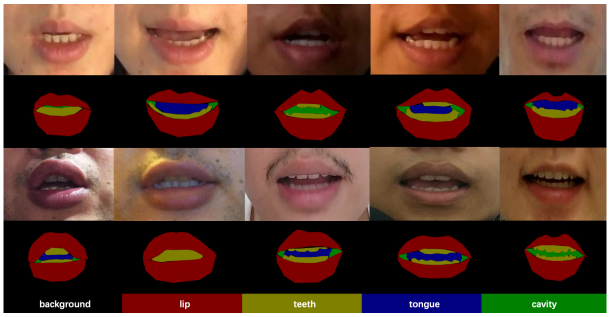


Fig. 5. Examples of lip images for training and the corresponding annotations. The proposed dataset has annotations of five categories, i.e., lip, teeth, tongue, inner cavity and background.

detection can be directly calculated from the segmentation annotation.

$$L_{edge} = -\sum_{i}^{hw} y_i \log y_i' + (1 - y_i)\log(1 - y_i') \qquad (17)$$

Besides, an auxiliary cross-entropy loss $L_{aux}$ is used to supervise the auxiliary head, which only utilizes feature maps extracted from the encoder and generate coarse segmentation results in the FCN structure. The auxliary head is introduced for deep supervision [39]–[41], which is beneficial for the model to converge during training.

$$L_{aux} = -\sum_{i}^{hw}\sum_{n}^{N} y_{i,n} \log y_{i,n}' \qquad (18)$$

## IV. Experiments and Discussions

### A. Experiment Setup

*1) Dataset:* Since existing lip segmentation and/or face parsing datasets [42], [43] only labeled lip pixels and lacked inner mouth component annotations, they cannot be directly applied in the fine-grained lip segmentation task. Hence, we build the Fine-grained Lip Region Segmentation[1] (FLRSeg in short) dataset to evaluate the fine-grained segmentation performance. 630 facial images were randomly selected from the Visual Speaker Authentication (VSA) dataset [44], which contains videos from 58 speakers talking in natural scenes captured by different mobile phones with various background and illumination conditions. For each facial image, the rough lip region was localized by Dlib[2] and used as the lip image for segmentation. All the lip images are annotated by LabelBee[3] with five semantics categories: background, lip, teeth, tongue, and inner cavity as shown in Fig. 5.

*2) Evaluation Metrics:* Evaluation Metrics: Three widely used metrics, including the mean IoU(mIoU), pixel accuracy (PA), and mean pixel accuracy (mPA) following [27], were adopted. These metrics are defined as follows.

(1) mIoU = $\frac{1}{k}\sum_i \frac{p_{ii}}{\sum_j p_{ij} + \sum_j p_{ji} - p_{ii}}$

TABLE I
Parameter search for the number of kernels in FLM and nodes in FGR over multi-scale features(from low-level to high-level)

| Kernels | Nodes | mIoU | PA | Params | FLOPs |
|---|---|---|---|---|---|
| 8/16/32 | 8/16/32 | 68.66% | 93.46% | 28.765M | 3.357G |
| 8/16/32 | 16/32/64 | 69.75% | 93.88% | 28.774M | 3.358G |
| 8/16/32 | 32/64/128 | 66.49% | 93.99% | 28.806M | 3.360G |
| 16/32/64 | 8/16/32 | 66.22% | 92.79% | 28.765M | 3.357G |
| 16/32/64 | 16/32/64 | **74.37%** | 94.53% | 28.774M | 3.358G |
| 16/32/64 | 32/64/128 | 73.41% | **94.64%** | 28.806M | 3.360G |
| 32/64/128 | 8/16/32 | 68.58% | 93.55% | 28.765M | 3.357G |
| 32/64/128 | 16/32/64 | 66.07% | 92.01% | 28.774M | 3.358G |
| 32/64/128 | 32/64/128 | 71.46% | 93.73% | 28.806M | 3.360G |

(2) PA = $\frac{\sum_i p_{ii}}{\sum_i \sum_j p_{ij}}$

(3) mPA = $\frac{1}{N}\sum_i \frac{p_{ii}}{\sum_j p_{ij}}$

where $p_{ij}$ means the number of pixels in the i-th class being classified as the j-th class.

As pixel accuracy pays more attention to classes with more pixels, PA in fine-grained lip image segmentation is usually dominated by the segmentation results of the lip and background classes and cannot accurately reflect the segmentation performance of the inner mouth components. Employing the average value of PA and IoU for different classes, i.e. mPA and mIoU, can alleviate this problem to some extent. Since mIoU is less sensitive to the scale of the mask compared to mPA [45], mIoU is the key evaluation metric in our experiment.

*3) Implementation Details:* The AdamW optimizer was used during training with an initial learning rate of 0.001 and a weight decay of 1e-5. The cosine learning rate policy was employed, where the initial learning rate decreased according to the cosine curve and reached the minimum value of 1e-6 at the end of training. The input of the network were first transformed to the size of 128*256 using crop and zero padding. The model was trained for 100 epochs with a batch size of 32 using two NVIDIA GeForce 3080 GPUs. The values for $k_{dice}$, $k_{aux}$ and $k_{edge}$ were empirically set to 1.0, 0.1 and 0.1, respectively .

In our experiments, we applied the following data augmentation processes during training: i) random color jitter, ii) random horizontal flipping, iii) random Gaussian noise, iv) random resizing and cropping. During evaluation, we used multi-scale inference with scales of 0.5, 1.0, 1.5, 2.0 and horizontal flip. The reported metrics were the average value obtained from three independent experiments initialized with different random seeds.

### B. Parameter Selections

To determine the optimal hyper-parameters, we performed a grid search for the number of membership functions (kernels) in FLM and the number of nodes in FGR. The experiment results are given in Table I. We used the thop[4] tool to calculate the FLOPs and number of network parameters.

It can be observed from the table that: i) the selection of the hyper-parameters in the proposed FLM and FGR has little impact on the total number of parameters and FLOPs of the

[1] https://github.com/YangLeiSX/FLRSeg
[2] http://dlib.net
[3] https://github.com/open-mmlab/labelbee
[4] https://github.com/Lyken17/pytorch-OpCounter

TABLE II
ABLATION STUDY OF PROPOSED FUZZY LEARNING MODULE(FLM) AND FUZZY GRAPH REASONING MODULE(FGR)

| Model | mIoU | PA | mPA | #Params(M) | GFLOPs |
|---|---|---|---|---|---|
| Baseline | 66.95% | 93.95% | 75.98% | 27.67(———) | 3.066(———) |
| +Rules | 70.17% | 93.52% | 76.60% | 28.57(+0.90) | 3.318(+0.252) |
| +Rules+Recons. | 72.20% | 93.49% | **83.03%** | 28.57(+0.90) | 3.318(+0.252) |
| +GR | 67.94% | 93.34% | 79.38% | 27.88(+0.21) | 3.122(+0.056) |
| +GR+Proj. | 69.34% | 90.68% | 77.36% | 27.86(+0.19) | 3.106(+0.040) |
| Proposed | **74.89%** | **94.36%** | 81.78% | 28.77(+1.10) | 3.358(+0.292) |

*Rule means with fuzzy rules in the proposed FLM, and Recons. means feature reconstruction.
* GR means with graph reasoning in the proposed FGR, and Proj. means fuzzy projection.

TABLE III
CLASS-WISE PIXEL ACCURACY IN ABLATION STUDY

| Model | bg | lip | teeth | tongue | cavity |
|---|---|---|---|---|---|
| Baseline | 97.28% | 91.78% | 67.36% | 82.09% | 41.38% |
| +Rules | 97.56% | 90.41% | 58.76% | 80.50% | 55.79% |
| +Rules+Recons. | **97.62%** | 89.21% | 61.42% | **88.23%** | **78.66%** |
| +GR | 97.22% | 89.86% | **69.12%** | 83.05% | 57.63% |
| +GR+Proj. | 87.22% | **98.11%** | 49.53% | 83.71% | 68.25% |
| Proposed | 94.20% | 96.86% | 60.48% | 84.02% | 73.34% |

TABLE IV
CLASS-WISE IOU IN ABLATION STUDY

| Model | bg | lip | teeth | tongue | cavity |
|---|---|---|---|---|---|
| Baseline | **92.83%** | 86.59% | 51.38% | 70.17% | 33.79% |
| +Rules | 91.94% | 85.01% | 52.82% | 72.13% | 48.99% |
| +Rules+Recons. | 91.67% | 85.22% | 51.03% | 72.84% | 60.24% |
| +GR | 91.80% | 85.46% | 45.40% | 72.11% | 44.95% |
| +GR+Proj. | 86.45% | 80.88% | 47.62% | 71.79% | 59.95% |
| Proposed | 92.65% | **87.57%** | **54.09%** | **73.38%** | **66.78%** |

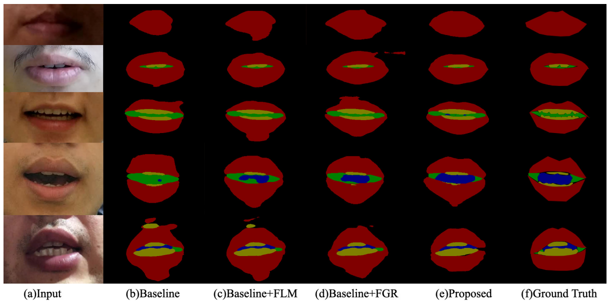


Fig. 6. Lip segmentation results with proposed networks.

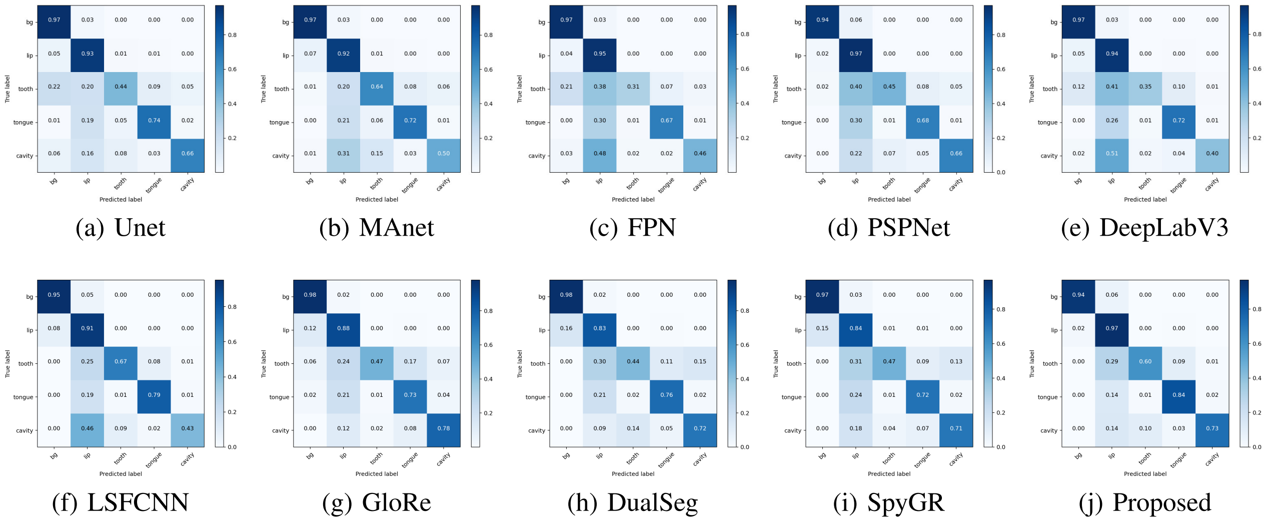
entire network; ii) the segmentation performance decreases when the number of kernels/nodes is set too small or too large. A limited number of kernels/nodes leads to insufficient modeling while too many abstract nodes results in overfitting of the underlying relationship among semantic parts and creates confusion for GCN. For subsequent experiments, we set the number of membership functions and the number of graph nodes to 16/32/64 and 16/32/64, respectively, as they provide the best segmentation performance with respect to mIoU.

### *C. Ablation Studies*

To investigate the effectiveness of the proposed FLM and FGR modules, ablation studies were carried out, and the experiment results are given in Table II.

From the table, it is observed that: i) Employing the fuzzy rules in FLM leads to an mIoU gain of 3.22% with a slight decrease in PA. This demonstrates that the fuzzy rules can better differentiate pixels of inner mouth components. In addition, the segmentation results are further improved w.r.t. mIoU by using feature map reconstruction from the output of the fuzzy rules. This is because compared with the fuzzy logic outputs, the distribution of the reconstructed feature map is more similar to that of the input feature map fed to FLM, making it easier for the network to learn the concatenated feature after fusion. ii) Reasoning through the constructed graph improves the segmentation performances in mIoU, demonstrating that modeling the relations between the components from a global perspective helps to localize the inner mouth components. The fuzzy projection in the proposed FGR further improves the mIoU with fewer parameters, but significantly decreases PA and mPA. The class-wise pixel accuracy and IoU are given in Table III and Table IV. It can be observed that the improvement of mIoU is mainly due to the improvements of teeth and inner cavity, and the decline of PA and mPA is primarily due to the decreased pixel accuracy of background pixels. The model with fuzzy projection tends to misclassify background pixels as lip, but this has less impact on downstream tasks. On the other hand, the better segmentation performance on inner mouth components is beneficial in tasks such as lipreading. iii) The proposed FLM and FGR work together to improve segmentation performance, and the network with both modules achieves a 7.94% improvement compared to the baseline. This is because FLM focuses on learning complex rules between pixel feature and semantic categories, while FGR focuses on modeling the relations between pixels with a global receptive field. The proposed model benefits from these two complementary capabilities.

Some segmentation results are shown in Fig. 6. From the figure, it can be seen that the baseline model struggles with blurry boundaries, particularly around the inner mouth component, due to the color overlapping problem under complex illumination conditions. With the blurred boundary between the inner mouth components and the lip region, false positive patches appear around the tongue and corners of the mouth. With FLM, false positive patches around the inner mouth components are reduced (Fig. 6c), which demonstrates that the model can effectively handle the uncertainties in images. With the proposed FGR, the model generates better result in the mouth region compared to the baseline model. This is



(a) Unet (b) MAnet (c) FPN (d) PSPNet (e) DeepLabV3

(f) LSFCNN (g) GloRe (h) DualSeg (i) SpyGR (j) Proposed

Fig. 7. Confusion matrix of existing models.

TABLE V
COMPARISON WITH EXISTED SEGMENTATION MODELS

| model | mIoU | PA | mPA | Params | FLOPs |
|---|---|---|---|---|---|
| Unet [33] | 66.72% | 94.23% | 74.90% | 24.44M | 3.944G |
| MAnet [25] | 65.21% | 93.80% | 74.98% | 31.78M | 4.197G |
| FPN [19] | 62.79% | 94.20% | 67.24% | 23.15M | 3.440G |
| PSPNet [23] | 67.41% | 93.89% | 74.15% | 1.52M | 1.190G |
| DeepLabV3 [46] | 62.47% | 93.88% | 67.67% | 26.01M | 13.678G |
| LSFCNN [27] | 68.56% | 92.49% | 75.06% | 28.37M | 3.262G |
| GloRe [34] | 64.51% | 92.56% | 76.70% | 27.65M | 3.060G |
| DualSeg [47] | 64.56% | 90.86% | 74.50% | 27.64M | 3.167G |
| SpyGR [48] | 63.87% | 90.48% | 74.12% | 21.44M | 2.584G |
| Proposed | **74.89%** | **94.36%** | **81.78%** | 28.77M | 3.358G |

because graph reasoning can model the relationships around semantics objects in the images and obtain a global reception field. With both FLM and FGR, the proposed model can better handle uncertainties and similar features around inner mouth components and achieve better segmentation performance.

### *D. Comparison with existing segmentation methods*

In order to comprehensively evaluate the segmentation performance of the proposed method, we compared it with several existing methods. Five of these methods, namely Unet [33], MANet [25], FPN [19], PSPNet [23] and DeepLabV3 [46], were trained on the FLRSeg Dataset using the open-source implementation called Segmentation Models in PyTorch (SMP)[5]. Three graph reasoning based methods GloRe [34], DualSeg [47] and SpyGR [48], and one lip segmentation network LSFCNN from our previous work [27], were also implemented and trained on FLRSeg. For a fair comparison, a pretrained ResNet-34 [49] network with the final pooling and fully connected layer removed was used to extract the fundamental image features for all the models. The experiment results are presented in Table V.

The results show that the proposed network can achieve a higher mIoU score compared to the other segmentation models with only a slight increase in overhead. It can be observed that existing semantic segmentation methods are not very effective on fine-grained segmentation. On the other hand, our proposed model with the FLM and FGR modules could model the complex semantic relationships in the image and deal with uncertainties, allowing it to achieve better mIoU and mPA scores compared with all the methods compared.

Table V shows that the improvement of the PA metric is not very significant. This is because most pixels in lip images belong to either lip or background, so these methods with lower mIoU scores can still get a relatively high PA with most pixels in the image correctly classified. The confusion matrix presented in Fig. 7 shows that the proposed neural network achieved better results, especially around inner mouth components. While all the methods achieved good classification accuracy over the lip (83%-97%) and background (94%-98%), most existing methods may fail over inner mouth components, especially pixels belonging to teeth. Errors mostly occurs by misclassifying inner mouth pixels as lip due to the indistinct contour between lip and inner mouth components. Our proposed network was able to reduce misclassification, especially over inner mouth components, and achieve higher accuracy.

Additional segmentation results are shown in Fig. 8. As can be seen, all models were able to accurately segment the lip region for the close mouth scenarios. But for lip images with open mouth, the proposed network outperformed the other methods, particularly around the corner of the mouth, teeth, and inner cavities, which cannot be handled by existing segmentation methods. In the second and the third row, motion blur can be seen in the lip area of the images extracted from speaker videos. The motion blur makes it difficult to determine the exact edge of the lips, resulting in misclassified pixels. The proposed network was able to reduce misclassification around the inner contour (shown in orange boxes) and outer contour of the lips (shown in blue boxes). In the fourth and fifth row of Fig. 8, we see that under variable illumination condition, existing segmentation methods may fail around the corner

[5] https://smp.readthedocs.io/en/latest/index.html

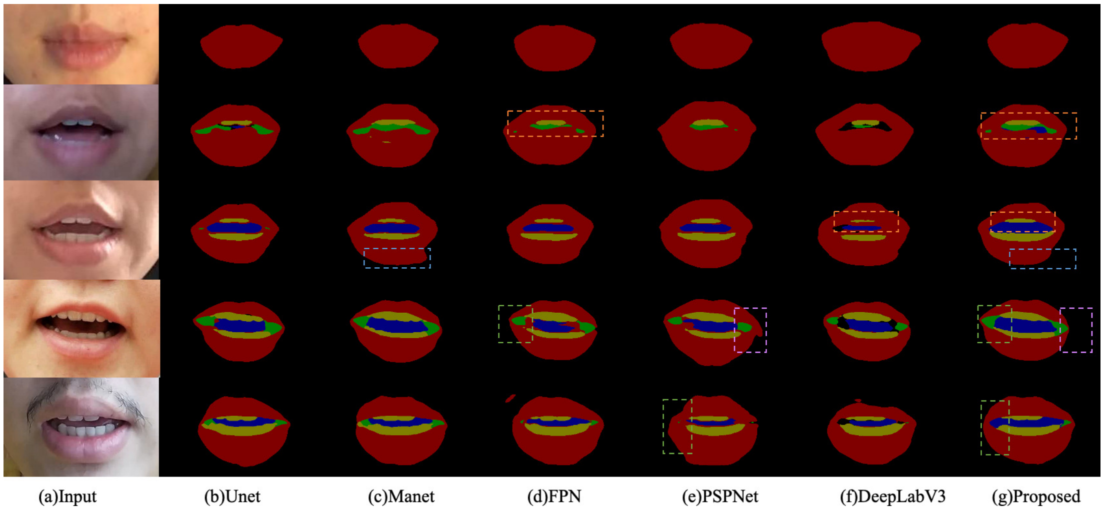


Fig. 8. Lip segmentation results on FLRSeg with existed segmentation methods and proposed network.

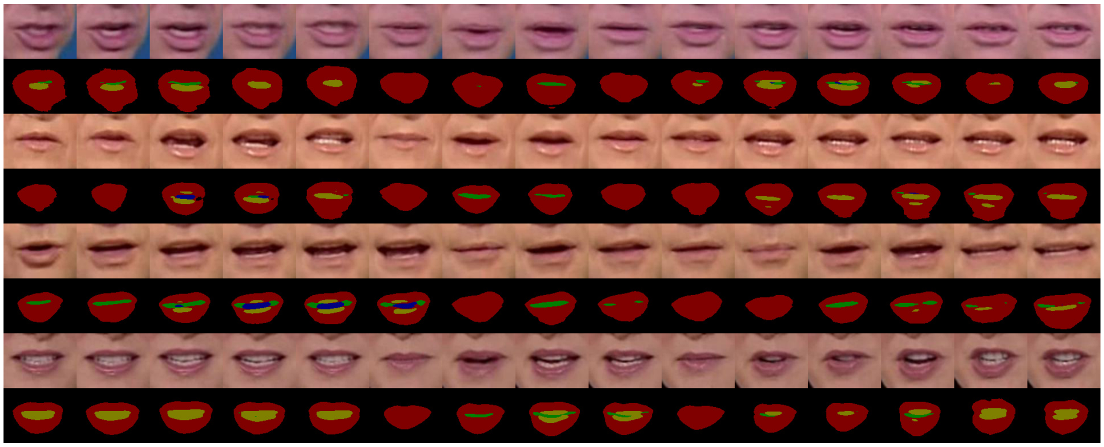


Fig. 9. Lip segmentation results on LRW.

of the mouth. In contrast, the proposed network was able to handle uncertainties and achieve better results, particularly around the tongue, inner cavity (shown in green boxes) and around the corner (shown in purple boxes).

### *E. Discussions*

*1) Cross-dataset Performance:* In order to comprehensively evaluate the transferability and scalability of the proposed model, the segmentations results on a large-scale dataset LRW [50], which is a word-level lip reading dataset consists of over 480,000 video clips, is shown in Fig. 9. As can be seen, the proposed models trained on the proposed FLRSeg dataset can generate fine-grained segmentation results for unseen speakers on both close mouth and open mouth scenarios, which is beneficial for future applications on downstream tasks.

*2) Computation Cost:* It can be observed from Table II that the main parameters and computation of the proposed network come from the encoder and decoder, which is used to extract multi-scale features and fuse them together. The proposed fuzzy learning module and fuzzy graph reasoning module are both lightweight. In out experiments, a pretrained ResNet-34 is adopted as the encoder for a fair comparison. However, the encoder can be replaced with an efficient neural network in an embedded system or for real-time applications. A comparison of different encoders is conducted, and the results are shown in Table VI. The reported FPS are calculated over a single RTX3080Ti with Intel Xeon 4210 CPU.

TABLE VI
COMPARISON WITH MODEL WITH DIFFERENT ENCODERS

| Encoder | mIoU | PA | Params | FLOPs | FPS |
|---|---|---|---|---|---|
| ResNet-34 [49] | 74.9% | 94.4% | 28.77M | 3.358G | 20.3 |
| ResNet-18 [49] | 66.2% | 93.3% | 18.67M | 2.146G | 23.9 |
| MobileNetV2 [51] | 71.1% | 93.0% | 9.19M | 1.297G | 22.9 |
| SqueezeNet [52] | 67.5% | 92.0% | 8.93M | 1.663G | 25.3 |
| ShuffleNetV2 [53] | 65.2% | 93.7% | 1.35M | 0.260G | 26.5 |

*3) Robust Evaluation:* To investigate the robustness of the proposed method to lighting, distortion and occlusion, we firstly collected some hard samples from the hold-out test set and applied the proposed model for prediction. The results are shown in the Fig. 10. It can be observed from the figure that the proposed neural network achieves good performance in scenarios that involve dynamic blurring and lip distortion. In situations where there is mustache in occlusion and strong

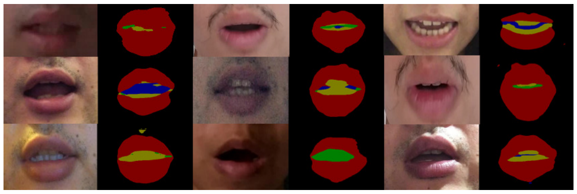


Fig. 10. Some hard cases in the test set of FLRSeg.

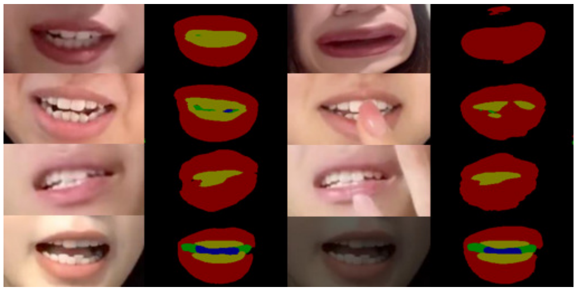


Fig. 11. Segmentation results for robustness evaluation.

noise, the proposed method still achieved robust results. The samples shown in the figure also demonstrate the robustness of our model under various lighting conditions, such as low light or side lighting.

Further, we collected additional samples outside the VSA dataset, which contain distortion, occlusion, and lighting change. The results are shown in Fig. 11, each row contains images from the same speaker and the corresponding segmentation results. It can be seen that the proposed method can handle lip distortion and change of brightness well. However, the model trained on FLRSeg dataset cannot well handle images with lip occlusion.

There are two main reasons for the above phenomenon. First, the FLRSeg dataset are extracted from the VSA dataset, which is a visual speaker dataset obtained in a natural setting that covers a variety of lighting conditions. However, lip images in the dataset are not occluded. Secondly, the proposed method utilizes fuzzy logic in neural network, where the neural network extracts discriminative features and the fuzzy learning module learns the complex semantic rules between pixel features and categories. Combining the ability of fuzzy systems to deal with uncertainties in images and annotations, our model is capable of handling ambiguity in images and exhibits robustness. However, the proposed model may not work well in some extreme scenarios, such as predicting errors around lip contour in images with low brightness and under occlusion. This can be improved by introducing additional annotated training samples in these extreme scenarios and appropriate data augmentation during training.

## V. Conclusion

In this paper, we proposed a new lip segmentation method based on fuzzy convolutional neural network with graph reasoning that can learn high-level semantics. The fuzzy learning module and the fuzzy graph reasoning module help the deep convolutional neural network to handle uncertainties around ambiguous boundaries, capture global information, and improve multi-class lip region segmentation. In addition, a fine-grained lip region dataset is released for multi-class segmentation studies. Our proposed approach achieved satisfactory performance on the test set, with 94.36% pixel accuracy and 74.89% mIoU. The experiment results have demonstrated that the proposed method can be applied in many lip-related applications to obtain accurate and robust lip region segmentation in natural scenes.

However, the proposed network cannot achieve satisfactory results in certain scenarios, specifically in cases of occlusion or extreme lighting conditions. This limitation can be attributed to the FLRSeg dataset, which is derived from the VSA dataset that was collected for speaker authentication and thus required unobstructed lip movements. In our future work, we will further improve the segmentation performance in various situations and explore the potential of leveraging the segmentation results in downstream tasks, such as lip reading and visual speaker authentication, to enhance performance and accelerate converge.

## References

[1] I. A. Matthews, T. F. Cootes, J. A. Bangham, S. J. Cox, and R. W. Harvey, *Extraction of Visual Features for Lipreading*, Std. 2, 2002.

[2] S. Wang, A. W. Liew, W. H. Lau, and S. H. Leung, “An automatic lipreading system for spoken digits with limited training data,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 12, pp. 1760–1765, dec 2008.

[3] H. E. Çetingül, Y. Yemez, E. Erzin, and A. M. Tekalp, “Discriminative analysis of lip motion features for speaker identification and speech-reading,” *IEEE Transactions on Image Processing*, vol. 15, no. 10, pp. 2879–2891, oct 2006.

[4] X. Liu and Y. Cheung, “Learning multi-boosted hmms for lip-password based speaker verification,” *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 2, pp. 233–246, feb 2014.

[5] C. Chan, B. Goswami, J. Kittler, and W. J. Christmas, “Local ordinal contrast pattern histograms for spatiotemporal, lip-based speaker authentication,” *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 2, pp. 602–612, apr 2012.

[6] G. Kodandaramaiah, M. Manjunatha, S. Jilani, M. Giriprasad, R. Kulkarni, and M. M. Rao, “Use of lip synchronization by hearing impaired using digital image processing for enhanced perception of speech,” in *2009 2nd International Conference on Computer, Control and Communication*. IEEE, 2009, pp. 1–7.

[7] N. Eveno, A. Caplier, and P.-Y. Coulon, “New color transformation for lips segmentation,” in *Fourth IEEE Workshop on Multimedia Signal Processing, MMSP 2001*, J. Dugelay and K. Rose, Eds. IEEE, 2001, pp. 3–8.

[8] M. Shemshaki and R. Amjadifard, “Lip segmentation using geometrical model of color distribution,” in *2011 7th Iranian Conference on Machine Vision and Image Processing*. IEEE, nov 2011.

[9] Y.-P. Guan, “Automatic extraction of lips based on multi-scale wavelet edge detection,” *IET Computer Vision*, vol. 2, no. 1, pp. 23–33, 2008.

[10] S. R. Banimahd and H. Ebrahimnezhad, “Lip segmentation using level set method: Fusing landmark edge distance and image information,” in *2010 20th International Conference on Pattern Recognition*. IEEE, aug 2010, pp. 2432–2435.

[11] C. Santiago, J. C. Nascimento, and J. S. Marques, “2d segmentation using a robust active shape model with the EM algorithm,” *IEEE Transactions on Image Processing*, vol. 24, no. 8, pp. 2592–2601, aug 2015.

[12] T. Cootes, G. Edwards, and C. Taylor, “Active appearance models,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681–685, jun 2001.

[13] Y. Cheung, M. Li, X. Cao, and X. You, *Lip Segmentation under MAP-MRF Framework with Automatic Selection of Local Observation Scale and Number of Segments*, Std. 8, aug 2014.

[14] A. W. Liew, S. H. Leung, and W. H. Lau, “Segmentation of color lip images by spatial fuzzy clustering,” *IEEE Transactions on Fuzzy Systems*, vol. 11, no. 4, pp. 542–549, aug 2003.

[15] S.-H. Leung, S.-L. Wang, and W.-H. Lau, "Lip image segmentation using fuzzy clustering incorporating an elliptic shape function," *IEEE transactions on image processing*, vol. 13, no. 1, pp. 51–62, 2004.

[16] S.-L. Wang, W.-H. Lau, A. W.-C. Liew, and S.-H. Leung, "Robust lip region segmentation for lip images with complex background," *Pattern Recognition*, vol. 40, no. 12, pp. 3481–3491, 2007.

[17] J. Fu, S. Wang, and X. Lin, "Robust lip region segmentation based on competitive FCM clustering," in *2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*. IEEE, nov 2016, pp. 1–8.

[18] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015*. IEEE Computer Society, jun 2015, pp. 3431–3440.

[19] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, jul 2017, pp. 936–944.

[20] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *4th International Conference on Learning Representations, ICLR 2016*, 2016.

[21] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE Computer Society, oct 2017, pp. 764–773.

[22] L. 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 Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, apr 2018.

[23] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*. IEEE Computer Society, jul 2017, pp. 6230–6239.

[24] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, jun 2018, pp. 7132–7141.

[25] T. Fan, G. Wang, Y. Li, and H. Wang, "MA-net: A multi-scale attention network for liver and tumor segmentation," *IEEE Access*, vol. 8, pp. 179 656–179 665, 2020.

[26] Z. Ju, X. Lin, F. Li, and S. Wang, "Lip segmentation with muti-scale features based on fully convolution network," in *2018 IEEE Third International Conference on Data Science in Cyberspace (DSC)*. IEEE, jun 2018, pp. 365–370.

[27] C. Guan, S. Wang, and A. W. Liew, "Lip image segmentation based on a fuzzy convolutional neural network," *IEEE Transactions on Fuzzy Systems*, vol. 28, no. 7, pp. 1242–1251, 2020.

[28] F. Ye, *Yu Yan Xue Gang Yao[Linguistics Outline]*. BEIJING BOOK CO. INC., 1997.

[29] R. Das, S. Sen, and U. Maulik, "A survey on fuzzy deep neural networks," *ACM Computing Surveys*, vol. 53, no. 3, pp. 54:1–54:25, may 2020.

[30] Y. Deng, Z. Ren, Y. Kong, F. Bao, and Q. Dai, "A hierarchical fused fuzzy deep neural network for data classification," *IEEE Transactions on Fuzzy Systems*, vol. 25, no. 4, pp. 1006–1012, aug 2017.

[31] Y. Zheng, W. Sheng, X. Sun, and S. Chen, "Airline passenger profiling based on fuzzy deep machine learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 12, pp. 2911–2923, dec 2017.

[32] S. Park, S. J. Lee, E. Weiss, and Y. Motai, "Intra- and inter-fractional variation prediction of lung tumors using fuzzy deep learning," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 4, pp. 1–12, 2016.

[33] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Lecture Notes in Computer Science*, ser. Lecture Notes in Computer Science, N. Navab, J. Hornegger, W. M. W. III, and A. F. Frangi, Eds., vol. 9351. Springer International Publishing, 2015, pp. 234–241.

[34] Y. Chen, M. Rohrbach, Z. Yan, S. Yan, J. Feng, and Y. Kalantidis, "Graph-based global reasoning networks," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Computer Vision Foundation / IEEE, jun 2019, pp. 433–442.

[35] Q. Li, Z. Han, and X.-M. Wu, "Deeper insights into graph convolutional networks for semi-supervised learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[36] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*. IEEE, 2016, pp. 565–571.

[37] T. Takikawa, D. Acuna, V. Jampani, and S. Fidler, "Gated-scnn: Gated shape cnns for semantic segmentation," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, oct 2019, pp. 5228–5237.

[38] Y. Yuan, J. Xie, X. Chen, and J. Wang, "Segfix: Model-agnostic boundary refinement for segmentation," in *Computer Vision – ECCV 2020*, ser. Lecture Notes in Computer Science, A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, Eds., vol. 12357. Springer, 2020, pp. 489–506.

[39] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, "Deeply-supervised nets," in *Artificial intelligence and statistics*. Pmlr, 2015, pp. 562–570.

[40] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[41] G. Zeng, X. Yang, J. Li, L. Yu, P.-A. Heng, and G. Zheng, "3d u-net with multi-level deep supervision: fully automatic segmentation of proximal femur in 3d mr images," in *Machine Learning in Medical Imaging: 8th International Workshop, MLMI 2017, Held in Conjunction with MICCAI 2017, Quebec City, QC, Canada, September 10, 2017, Proceedings 8*. Springer, 2017, pp. 274–282.

[42] V. Le, J. Brandt, Z. Lin, L. D. Bourdev, and T. S. Huang, "Interactive facial feature localization," in *Computer Vision – ECCV 2012*, ser. Lecture Notes in Computer Science, A. W. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, Eds., vol. 7574. Springer, 2012, pp. 679–692.

[43] C. Lee, Z. Liu, L. Wu, and P. Luo, "Maskgan: Towards diverse and interactive facial image manipulation," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020*. Computer Vision Foundation / IEEE, jun 2020, pp. 5548–5557.

[44] J. Sun, S. Wang, and Q. Zhang, "Visual speaker authentication by a CNN-based scheme with discriminative segment analysis," in *Communications in Computer and Information Science*, ser. Communications in Computer and Information Science, T. Gedeon, K. W. Wong, and M. Lee, Eds., vol. 1142. Springer, 2019, pp. 159–167.

[45] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. D. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Computer Vision Foundation / IEEE, jun 2019, pp. 658–666.

[46] L. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *CoRR*, vol. abs/1706.05587, 2017.

[47] L. Zhang, X. Li, A. Arnab, K. Yang, Y. Tong, and P. H. S. Torr, "Dual graph convolutional network for semantic segmentation," in *30th British Machine Vision Conference 2019, BMVC 2019, Cardiff, UK, September 9-12, 2019*. BMVA Press, 2019, p. 254.

[48] X. Li, Y. Yang, Q. Zhao, T. Shen, Z. Lin, and H. Liu, "Spatial pyramid based graph reasoning for semantic segmentation," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Computer Vision Foundation / IEEE, jun 2020, pp. 8947–8956.

[49] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[50] J. S. Chung and A. Zisserman, "Lip reading in the wild," in *Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13*. Springer, 2017, pp. 87–103.

[51] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[52] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size," *arXiv preprint arXiv:1602.07360*, 2016.

[53] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient cnn architecture design," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 116–131.