Lip Image Segmentation Based on a Fuzzy Convolutional Neural Network

Lip image segmentation based on a fuzzy convolutional neural network

Cheng Guan, Shilin Wang*, *Senior Member, IEEE* and Alan Wee-Chung Liew, *Senior Member, IEEE*

***Abstract*— Research has shown that the human lip and its movements are a rich source of information related to speech content and speaker's identity. Lip image segmentation, as a fundamental step in many lip-reading and visual speaker authentication systems, is of vital importance. Because of variations in lip color, lighting conditions and especially the complex appearance of an open mouth, accurate lip region segmentation is still a challenging task. To address this problem, this paper proposes a new fuzzy deep neural network having an architecture that integrates fuzzy units and traditional convolutional units. The convolutional units are used to extract discriminative features at different scales to provide comprehensive information for pixel-level lip segmentation. The fuzzy logic modules are employed to handle various kinds of uncertainties and to provide a more robust segmentation result. An end-to-end training scheme is then used to learn the optimal parameters for both the fuzzy and the convolutional units. A dataset containing more than 48,000 images of various speakers, under different lighting conditions, was used to evaluate lip segmentation performance. According to the experimental results, the proposed method achieves state-of-the-art performance when compared with other algorithms.**

***Index Terms*— fuzzy neural networks, convolutional neural network, lip region segmentation**

I. Introduction

Lip image processing has attracted wide-spread research interest in recent years for its wide application in automatic visual speech recognition [1-2], visual speaker authentication [3-5], lip synchronization for facial animation [6], etc. Lip region segmentation, which is also referred to as lip segmentation, is the first and most crucial step in various lip-related applications [7].

In past decades, many researchers have proposed various lip segmentation approaches [8-22] which can basically be divided into three categories depending on the information source they exploit, i.e. color-based, edge-based or spatial information guided approaches. The color-based approaches [8-9] normally detect lip pixels by a preset color filter which is able to differentiate lip and non-lip pixels in a specific color space. The color-based approaches can obtain good segmentation result for lip images where there is a high color contrast. Therefore, in some early lip-based applications, make-up with lipstick was required.

The edge-based lip segmentation approaches (also referred to as gradient-based approaches) [10-11] aim to segment the lip region by detecting the lip-background boundary based on the grayscale or color edge information. Owing to the fact that most lip images are low contrast in nature, the edge strength in some background regions (such as mustache, teeth, tongue, moles, etc.) could be greater than that in the lip-background boundary. This gives rise to false edges that significantly degrade segmentation performance. In order to solve this problem, many researchers utilised various lip models, which introduced prior information for valid lip shapes and appearance. The active contour model (also called Snakes) [12,13], the Active Shape Model (ASM) [14] and the Active Appearance Model (AAM) [15] are the most widely used lip models. Lip models can provide geometric constraints for the final lip shape and reduce the influence caused by false edges to some extent. However, research [16] has demonstrated that the lip model based approaches depend on good initialization and are still sensitive to the noise boundary caused by various background regions such as mustache, teeth, etc.

Spatial information guided approaches [16-22] take the spatial location information into account to enhance the segmentation performance. Two kinds of spatial information have been widely utilized: local spatial information [16-19]; and global spatial information [20-22]. Local spatial information refers to the intrinsic connections between neighboring pixels and assumes that adjacent pixels are more likely to be in the same class (lip or background). Markov Random Field (MRF) based approaches are the typical local spatial information guided lip region segmentation methods [16-18]. Lievin et al. [17] and Zhang et al. [18] proposed two pioneering lip segmentation approaches based on MRF. Both color and edge information were considered in [18], and with the local spatial constraints, the lip segmentation performance was improved. In 2014, Cheung et al. proposed a hierarchical multi-scale MRF model for lip segmentation [16]. In this approach, the number of segments was automatically estimated

Cheng Guan and Shilin Wang are with the School of Electronic Information and Electrical Engineering, Shanghai Jiaotong University, 200240, Shanghai, China. (e-mail: Gclalaboo@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 is fully supported by NSFC Fund (No. 61771310).

by examining the lip images at various scales. Fuzzy clustering based methods are another widely used lip segmentation approach exploiting the spatial information [19-22]. Fuzzy C-means (FCM) clustering provides a flexible architecture, which can take both the color and spatial information into consideration. In [19], Liew et al. provided a dissimilarity measure integrating the local neighborhood information with the color information to enhance the segmentation performance. The above local spatial information guided approaches perform well on images with "pepper" noise; however, they often produce patches outside and holes inside the segmented lip region.

Global spatial information usually refers to the approximate location of the mouth region. In previous works of our group [20-22], we proposed an FCM based segmentation algorithm exploiting the global spatial information. Such information was seamlessly integrated in the dissimilarity measure to enhance/reduce the lip class probability for pixels inside/outside the mouth region. Compared with the local spatial information, the global spatial information is more useful for lip segmentation if the approximate mouth region is estimated accurately. One major disadvantage of these approaches is that global information does not effectively differentiate possible inner mouth components, especially the tongue region.

Generally speaking, the existing sophisticated lip segmentation approaches achieve good segmentation results to some extent in a closed mouth situation. However, for various lip images with open mouths, most of the above approaches do not perform well with the major difficulty being in segmenting the inner mouth components. This issue arises because the color and spatial location of the inner mouth components may be very similar to those of the lip region under various kinds of lighting conditions and thus it is very difficult to design universal rules or features to differentiate them.

With the rapid development of suitable hardware and software, deep neural networks have shown superior performance in image processing and computer vision [23-25], which provides a promising direction for solving the difficulties in lip segmentation. In recent work by our group [26], we have designed the Lip Segmentation Network (LSN) for examining the characteristics of the lip images. LSN adopts the classical fully convolutional network (FCN) [27] structure and integrates additional information from neighboring frames in a lip image sequence to reduce adverse effects caused by image noise and in-exact annotations. However, when analyzing these segmentation results, many misclassified pixels were observed in the inner mouth regions.

In order to solve the problems caused by the appearance of inner mouth components, we adopt fuzzy learning to achieve a discriminative feature representation for lip segmentation. Recent research [28-30] has applied fuzzy logic to machine learning fields.

Fuzzy learning can overcome various kinds of uncertainties in both raw lip images and corresponding annotations. Meanwhile, it is worth noting that according to human experience, some useful characteristics or rules for lip segmentation are fuzzy in nature, e.g. i) lip outer contour and inner contour are of an elliptic-like shape; ii) lip pixels are of similar color; and iii) pixels of teeth, gum and tongue are located inside the mouth region and gathered together to form patches with specific shapes, etc. In view of this, a new fuzzy logic feature representation is proposed in this paper for lip region segmentation. Inspired by [31], the fuzzy learning module is seamlessly integrated into the proposed deep neural network and thus the requirements of end-to-end learning can be satisfied. Considering both the segmentation performance and computational complexity, the deterministic part of the proposed network is composed of a series of convolutional layers. Hence, our method is referred to as the Lip Segmentation with Fuzzy Convolutional Neural Network (LSFCNN).

The major contributions of LSFCNN are three-fold: i) A novel network structure that integrates fuzzy logic into the convolutional neural network for lip segmentation is proposed. To our best knowledge, this is the first attempt to adopt fuzzy neural network in lip segmentation; ii) The proposed fuzzy learning module can provide interpretable results which helps us to understand the underlying mechanism of the segmentation network; iii) In the training stage, a new loss function is proposed which can guide the network to provide accurate segmentation results in an efficient manner. In addition, end-to-end training is also implemented.

The rest of the paper is organized as follows. Section II shows the difficulties and challenges in lip image segmentation, especially in the open mouth situation. Section III describes the detail architecture of LSFCNN, including the model structure, fuzzy logic network, the loss function, and other implementation details. In section IV, the dataset and the experimental results are presented as well as the ablation study that evaluates the necessity of each component of LSFCNN. Finally, Section V contains a conclusion and summary of the results obtained.

II. Challenges in lip segmentation

A. Problem Description in Lip segmentation

Segmenting the lip region from lip images taken in uncontrolled environments is a challenging problem especially when the mouth is open and various inner-mouth components (i.e. teeth, tongue, etc.) are visible [22]. Fig.1 shows an example of a lip image with an open mouth and its corresponding characteristics. The following issues can be observed from this figure. First, the color of the lip pixels and non-lip pixels significantly overlap each other (as shown in Fig.1b). Due to variations in illumination, shadow, etc., the colors of the lip pixels are not uniform. With the tongue and gums visible, pixels in these two regions have colors very similar to lip pixels. Such color similarity causes unsolvable problems for color-based lip segmentation algorithms. Second, there are many edges located inside and outside the mouth region (as shown in Fig.1c) and there is no obvious gradient change along the lip-background boundary when compared with the rest of the image (as shown in Fig.1d). In this case, the edge based and gradient based algorithms become confused and are misled by

these false edges and are unable to achieve good segmentation results. Third, the pixels of inner mouth components, such as teeth, tongue, gums, etc., have spatial locations that are close to the lip pixels. Hence, spatial information does not provide useful cues for differentiating between lip pixels and inner mouth pixels. As a consequence, color, gradient and spatial information alone cannot be used to segment the lip region accurately in an open-mouth scenario. How to discover the intrinsic rules in lip region segmentation and comprehensively exploit the above information remains a challenging question.

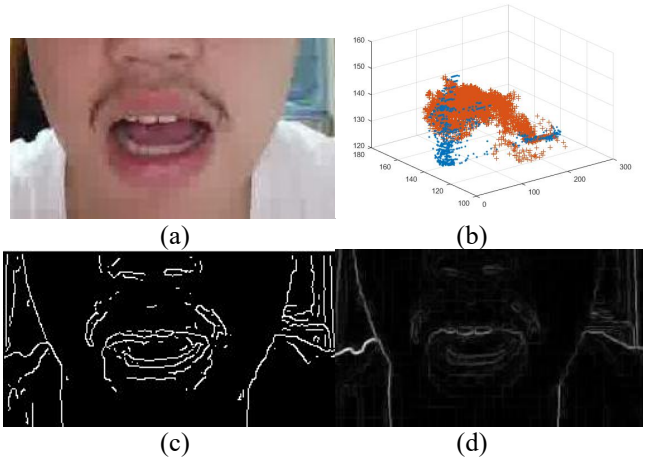

Fig. 1. (a) The original image, (b) (c) and (d) are the color distribution in RGB color space (red pixels denote the lip pixels and blue ones denote the non-lip pixels), the edge image by the Canny operator, and the intensity gradient map of (a), respectively.

B. Lip Segmentation with Fuzzy Convolutional Neural Network (LSFCNN)

To address the difficulties described above, a new neural network structure that combines the merits of deep convolutional neural network and fuzzy learning is proposed. The underlying rationale of using such a structure for lip segmentation is as follows.

Fig.2 Pixels that are difficult to classify even for human

According to the discussions in Section IIA, using color, spatial, and gradient information individually cannot achieve good segmentation results. In addition, simple handcrafted rules combining the above information (e.g. in [21] where the spatial and color information is integrated in the dissimilarity function) also fail to segment lip regions with open-mouths accurately and robustly. Instead, robust and accurate lip segmentation requires a series of complex rules or features that can better exploit the color, spatial and gradient information and their intrinsic relationship. Hence, a deep convolutional neural network structure is adopted in our approach as this can learn complex features from the training samples with lip-pixel annotations. As shown in [16], lip segmentation at different resolutions gives quite different results. Therefore, a multi-scale, hierarchical network structure was designed where the local context information and the global view information are represented in the feature maps at different resolutions. Low level feature maps with small receptive fields depict the local contexts around a pixel. High level feature maps with large receptive fields considered the global location and semantic relationship among various kinds of objects such as the lip, teeth, skin, mustache, etc. Taking all these feature maps into consideration, a comprehensive view of the lip image can be obtained.

On the other hand, the definition of a lip pixel is in fact a fuzzy concept. As shown in Fig.2, for those pixels in the red zone, even humans cannot judge whether they are lip pixels or not and different people may annotate them differently. To handle such uncertainty in annotation, fuzzy learning is adopted in our algorithm and the fuzzy learning modules are seamlessly integrated into the neural network. Each fuzzy learning module follows a feature map at a specific scale and aims to build a connection between the features and the segmentation result. By considering the outputs of the fuzzy learning modules at various levels, a robust and accurate lip segmentation result can be achieved.

The multi-scale deep convolutional feature maps and the fuzzy learning modules are the two key ingredients for robust lip region segmentation and they work together to overcome the difficulties in lip segmentation, especially in the open-mouth scenario. Without the multi-scale feature maps, the fuzzy learning component cannot handle the variations caused by different illuminations, head pose, mouth shape and appearance, etc. Without the fuzzy learning module, the deterministic deep convolutional neural network cannot handle the uncertainties in lip pixel annotation. Hence, in the proposed network structure, these two components are integrated together to provide a robust segmentation result.

III. THE PROPOSED METHOD

The overall network structure of the Lip Segmentation with Fuzzy Convolutional Neural Network (LSFCNN) is given in Fig.3. As can be seen, LSFCNN is composed of two parts with different functionalities, i.e. the deep convolutional subnet for multi-scale image feature extraction and the fuzzy learning module to extract high-level semantic features in consideration of various uncertainties. The fuzzy learning modules are seamlessly integrated into the network and each module handles the feature maps at a particular image scale. The parameters for the deep convolutional subnet and the fuzzy learning modules are optimized together by end-to-end learning from the training samples with manually annotated lip regions. Details of LSFCNN are presented in the following subsections.

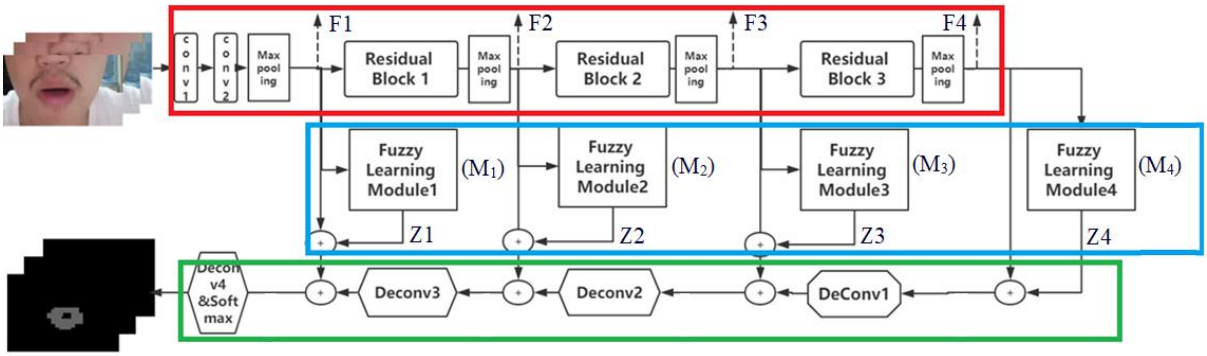

Fig. 3 Network structure of LSFCNN

A. *The Deep Convolutional Subnet*

The network structure of the deep convolutional subnet is similar to that of the fully convolutional network (FCN) [27], and includes the hierarchical feature extraction pipeline extracting from the low-level, fine image features to the high-level, coarse image features and the hierarchical feature map fusion pipeline integrating all the feature maps at various scale to a unified feature map for prediction.

In the feature extraction pipeline, inspired by Resnet [32], residual blocks as shown in Fig. 4 are adopted to extract the discriminative features at each scale. As shown in Fig. 3, the input images are first passed through two convolutional layers, followed by a max-pooling layer with a stride of 2 to generate the finest feature map F1 (of half the image size in both the horizontal and vertical direction). Then, a residual block followed by a max-pooling layer with a stride of 2 is applied to extract the comprehensive features in the corresponding resolution. This procedure is repeated three times and the feature maps F2 to F4 are extracted to represent the image characteristics at various resolutions.

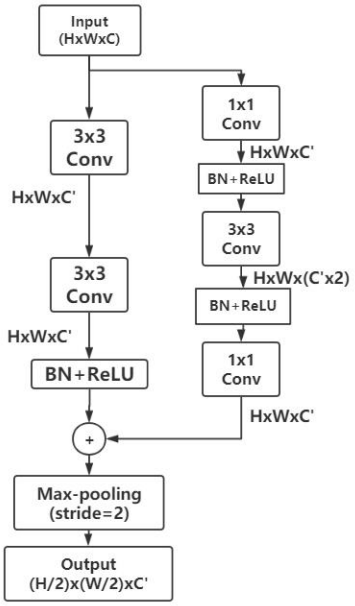

Fig. 4. Detailed structure of residual blocks, where H, W, C and C' denote the height, width of the feature map and the number of channels in the input and output feature map, respectively.

In the feature fusion pipeline, feature maps at various scales are fused together by the addition operation. To ensure that the size of the feature maps at different levels are the same, the deconvolution and upsampling operations are adopted, which runs as follows (Fig. 5): i) zeros are inserted into the gap between adjacent points; and ii) a 2×2 convolutional filter is adopted to upsample the previous feature map. Finally, the finest fused feature map is passed through a deconvolution and upsampling layer and a softmax layer to generate the final segmentation result.

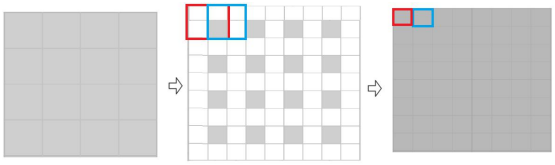

Fig.5 Structure of the deconvolution and upsampling layer.

B. *The Fuzzy Learning Module*

The objective of the fuzzy learning module is to learn the intrinsic, complex rules between the feature map and the corresponding segmentation result. For each feature map from F1 to F4, a specific fuzzy learning module with the same structure as shown in Fig.6 is adopted. Note that the fuzzy learning module processes only the feature maps rather than the raw image data to avoid confusions caused by variations in lighting, translation, rotation, etc.

Let F be the input feature map with a size of $H\times W\times C$, where H and W denote the height and width of the feature map, respectively and C is the number of channels. As shown in Fig.6, for a specific channel c, M membership functions are applied to each feature point in the channel. It is noted that M is kept the same for each channel of the feature map and can vary among different input feature maps. Each membership function assigns a fuzzy linguistic term label for the feature points and all the membership functions are in the form of the Gaussian function given in (1).

$$Z_{x,y,k,c} = e^{-\left(\frac{F_{x,y,c}-\mu_{k,c}}{\sigma_{k,c}}\right)^2}, \quad x = 1..W, y = 1..H, k = 1..M \qquad (1)$$

where (x,y) is the coordinate of the feature point $F_{x,y,c}$ in channel c, $\mu_{k,c}$ and $\sigma_{k,c}$ are the mean and standard deviation of the k-th Gaussian membership function and $Z_{x,y,k,c}$ represents the k-th output fuzzy linguistic term label of the feature point (x,y) in channel c. Note that $\mu_{k,c}$ and $\sigma_{k,c}$ $(1 \le k \le M, 1 \le c \le$

C) are trainable features which will be optimized during training. Similar to [31], the "AND" fuzzy logic is applied to all the memberships of the feature point and the final fuzzy degree of $F_{x,y,c}$ (which is denoted by $Z_{x,y,c}$) is obtained by

$$Z_{x,y,c} = \prod_{k=1}^{M} Z_{x,y,k,c} \quad (2)$$

As part of the network, the fuzzy learning module can be regarded as a fuzzy layer described by a number of parameters from the Gaussian membership functions. Note that the parameters $\mu_{k,c}$ and $\sigma_{k,c}$ ($k = 1..M$) remain unchanged in the same channel and can vary among different channels. This is because feature points in the same channel are extracted by the same convolutional kernel and will have similar characteristics; while feature points in a different channel are obtained using a different convolutional kernel and have different characteristics. Hence, the number of parameters in each fuzzy learning module can be calculated as $M \times C \times 2$, which is comparable to that in a single convolutional layer. Furthermore, to implement the fuzzy module, the number of membership functions, i.e. M, is the only hyper-parameter to be preset.

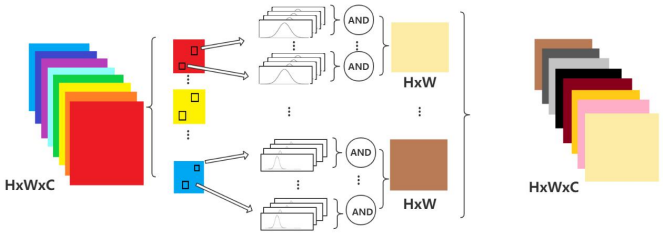

Fig. 6. Structure of the fuzzy learning module.

Finally, the output of the fuzzy learning module is the fuzzy degree tensor $Z=\{Z_{x,y,c}\}$ ($x = 1..W, y = 1..H, c = 1..C$), which has the same size as the input feature map tensor F. Before integration, the feature maps F1 to F4 and the fuzzy degree tensor Z1 to Z4 are processed by batch normalization (BN) to constrain their dynamic range. Then, a simple addition operation is adopted to integrate the fuzzy logic information.

C. Loss Function Formulation

By classifying the pixels into either the lip or non-lip class, a segmentation of the lip can be obtained. In most cases, the number of non-lip pixels is much larger than the number of lip pixels. However, traditional cross entropy loss cannot deal with such severe class imbalance problem, so in [33] Lin et al. proposed the focal loss to ease this problem, which is formulated as follows in our application,

$$FL = -\sum_{(x,y)\in R_{lip}} (1-p_{x,y})^{\beta} \log p_{x,y} - \sum_{(x,y)\in R_{non-lip}} p_{x,y}^{\beta} \log(1-p_{x,y}) \quad (3)$$

where $p_{x,y}$ is the estimated lip class probability of the pixel with the coordinate (x,y), R_{lip} and $R_{non\text{-}lip}$ refer to the coordinate sets belonging to lip and non-lip classes, respectively, and β is the hyperparameter with a positive value. By multiplying the factor of $(1-p_{x,y})^{\beta}$ for the lip pixels, the contribution of the easily classified pixels with a large lip class probability are weakened and vice versa. Hence, the difficult samples are emphasized in each iteration which eases the class imbalance problem and speeds up optimization. Inspired by [33], the following loss function is proposed according to the characteristics of the lip segmentation task, and is referred to as the lip loss ($Loss_{lip}$), i.e.,

$$Loss_{lip} = -\sum_{(x,y)\in R_{lip}} \left(1 + e^{\alpha + (1-p_{x,y})^{\beta}}\right) \log p_{x,y} - \sum_{(x,y)\in R_{non-lip}} \left(1 + e^{\alpha + p_{x,y}^{\beta}}\right) \log\left(1 - p_{x,y}\right) \quad (4)$$

where α, β are hyper-parameters with positive values. The differences between the proposed loss function, the classical cross-entropy (CE) loss function, and the focal loss function are shown in Fig. 7 for the lip class pixels. From the figure, it is observed that the function curves of the proposed lip loss and the focal loss are both less than that of the CE loss when lip probability is relatively high, i.e. when > 0.3, which helps the network to focus on those difficult samples (e.g. the moles with lip-like color in the background region, etc.) and the easily classified pixels dominating the background region will not be emphasized. Moreover, the proposed loss function decreases less rapidly compared to the CE loss when the lip class probability is low and has a higher value for larger probability values compared to focal loss [33]. Our loss function is designed based on the fact that: i) under the open mouth scenario, the pixels in the tongue, gums, etc. are of a very high lip-class probability at the start of the optimization stage. Hence, misclassifying such samples is given a very high penalty and the network is forced to memorize the characteristics of these challenging samples; ii) it is observed in our experiments that the lip class probabilities for a background pixel varies greatly during the early optimization stage. Hence, giving a relatively higher loss value for the background pixels with lip probabilities greater than, say 0.7, rather than ignoring them will help improve the robustness and stability during network optimization. The effectiveness of the proposed lip loss will be further demonstrated in Section 4.

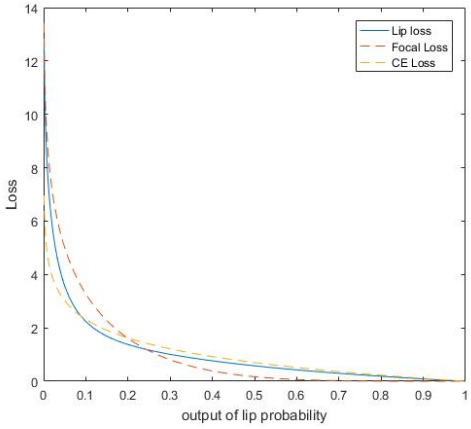

Fig 7. Function curves of the lip loss, CE loss, and focal loss for the lip class pixels. Note that the curves are linearly normalized where the area under curve is equal to one.

D. Implementation Procedures

As stated in the previous subsections, LSFCNN is composed of a deep convolutional subnet and fuzzy learning modules related to the feature maps at various scales. An end-to-end training scheme is used to obtain the optimum

network parameters by minimizing the loss function defined in Eqn. 4. We implemented LSFCNN on Pytorch 0.4 as follows.
i) Initialization: For the convolution and deconvolution layers in the deep convolutional subnet, the weights are initialized using a zero-mean Gaussian distribution with a standard deviation of $\sqrt{2/n}$, where n represents the number of weights in the layer. The bias in each layer is initialized as zero. For each fuzzy learning module, all the parameter pairs $\{\mu, \sigma\}$ in the membership functions are initialized using zero-mean, unit variance Gaussian distributions.
ii) Training: The batch size is set as 10. The stochastic gradient descent (SGD) [34] is adopted to optimize the network. Note that the training data is shuffled to guarantee that the lip images in one batch belong to different people. The learning rate is set to 10^{-6} and the momentum is set as 0.95.

TABLE I DETAILED STRUCTURE OF THE PROPOSED NETWORK, WHERE C AND C' ARE THE NUMBER OF CHANNELS OF THE INPUT AND OUTPUT FEATURE MAP, RESPECTIVELY.

Layer	kernels	C	C'	Parameters	With BN layer
Conv1	3X3	3	32	864	√
ReLU	-	-	-	0	
Conv2	3X3	32	64	18,432	√
ReLU	-	-	-	0	
Max pooling	-	-	-	0	
Res- Block1	3X3 & 1X1	64	128	557,056	√
ReLU	-	-	-	0	
Max pooling	-	-	-	0	
Res- Block2	3X3 & 1X1	128	256	2,228,224	√
ReLU	-	-	-	0	
Max pooling	-	-	-	0	
Res- Block3	3X3 & 1X1	256	512	8,912,896	√
ReLU	-	-	-	0	
Max pooling	-	-	-	0	
Fuzzy module 1	-	64	64	128	
Fuzzy module 2	-	128	128	512	
Fuzzy module 3	-	256	256	4096	
Fuzzy module 4	-	512	512	8192	
Deconv1	2X2	256	256	524288	√
Deconv2	2X2	128	128	262144	√
Deconv3	2X2	64	64	131072	√
Deconv4	2X2	2	2	512	√

iii) Inference: A test image is forwarded into the trained LSFCNN to obtain the segmentation result.

Details of the network hyper parameters are summarized in Table I and we provide the source code in "https://github.com/sjtuGC/LSFCNN". As it has been demonstrated that the increase in the depth of neural networks can reduce the accuracy [32], we used three residual blocks to achieve a good segmentation performance while maintaining lightweight. The batch normalization layers can speed up the training process and improve the convergence [35]. The fusion method in the residual block, multi-level feature maps, and fuzzy module is simply the addition operation.

IV. EXPERIMENT AND DISCUSSION

To evaluate the performance of the proposed network, a dataset containing 49 people was employed. Each person was asked to pronounce twenty phrases containing four digits in random. Each phrase consists of 54 frames. Note that for each speaker, the lip images were captured using his/her own cellphones under at least four different circumstances. By eliminating the low quality images, there are over 48,000 lip images in the dataset and some sample lip images are shown in Fig.8.

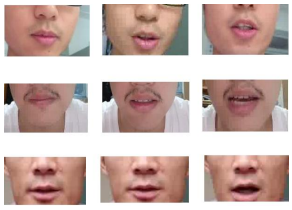

Fig.8 Sample images in our lip dataset

A. Experiment Setup

Since our approach uses supervised learning, the lip region of all the lip images in the dataset were manually annotated. The lip images of thirty speakers randomly selected from the dataset were adopted as the training data and those of the remaining nineteen speakers were adopted for testing. Four conventional metrics in [27] were adopted to evaluate the segmentation performance, i.e., the pixel-level classification accuracy, the mean pixel accuracy, the mean intersection over union (IU in short) and the frequency weighted IU. These metrics are formulated as follows,

1) pixel accuracy= $\frac{c_{lip}+c_{non-lip}}{n_{lip}+n_{non-lip}}$
2) mean accuracy= $\frac{1}{2}(\frac{c_{lip}}{n_{lip}}+\frac{c_{non-lip}}{n_{non-lip}})$
3) mean IU= $\frac{1}{2}(\frac{c_{lip}}{n_{lip}+n_{non-lip}-c_{non-lip}}+\frac{c_{non-lip}}{n_{non-lip}+n_{lip}-c_{lip}})$
4) frequency weighted IU= $\frac{1}{n_{lip}+n_{non-lip}}(\frac{n_{lip}c_{lip}}{n_{lip}+n_{non-lip}-c_{non-lip}}+\frac{n_{non-lip}c_{non-lip}}{n_{non-lip}+n_{lip}-c_{lip}})$

where c_{lip}, $c_{non-lip}$ denote the number of lip/non-lip pixels correctly classified as the lip/non-lip class, respectively; n_{lip}, $n_{non-lip}$ denote the total number of lip/non-lip pixels, respectively.

To overcome the variations caused by different lighting conditions, speaker's pose, distance towards the camera, etc., data augmentation was adopted. In our experiments, the following random processing steps had been performed as data augmentation: i) random flipping and cropping; ii) random noise on the brightness component; iii) random noise on the contrast component. After data augmentation, all the images were resized to 224×112 and then fed to the network.

B. Effectiveness of the Loss Function

TABLE II SEGMENTATION ACCURACIES IN % USING VARIOUS SELECTIONS OF HYPER PARAMETERS.

Accuracy in %	α=0.01	α=0.1	α=2	α=5	α=10
β=0.1	97.8	97.9	97.9	98.0	97.8
β=1	97.5	98.4	**98.5**	98.2	97.0
β=2	97.0	97.9	98.1	97.2	96.3
β=5	97.9	97.7	97.5	95.9	96.1
β=10	97.8	97.3	96.3	96.0	94.9

There are two hyper parameters α and β, in the proposed loss function that determine the shape and steepness of the loss curves. To select the optimal hyper parameters, a number of

parameters were investigated, i.e. $\alpha \in \{0.01,0.1,2,5,10\}$ and $\beta \in \{0.1,1,2,5,10\}$ and the pixel accuracy was adopted as the evaluation criterion. Table II shows the segmentation performance with various hyper parameter settings. From the table, it is observed that the highest pixel accuracy is achieved when α=2 and β=1. In addition, the accuracies obtained by the hyper parameters in the range of $1 \leq \alpha \leq 2$ and $0.1 \leq \beta \leq 2$ are similar, which demonstrates that the proposed loss function is insensitive to the hyper parameter settings.

With the optimal α and β, the lip segmentation performance during training is shown in Fig. 9. The classical Cross-Entropy (CE) loss and the Focal Loss (FL) [33] were also analyzed for comparison. From the figure, it is observed that within a small number of iterations (e.g. 3,000), the proposed lip loss can achieve an acceptable segmentation performance (with a mean IU over 0.8). In addition, the mean IU curves obtained for lip loss are usually above those obtained by CE loss and FL loss, which demonstrates that the proposed loss function speeds up network convergence.

As discussed in Section IIIC, compared with the focal loss, the proposed lip loss assigns a higher penalty for the background pixels especially for those with a non-lip class probability higher than 0.7. Hence, the misclassified background pixels is more quickly corrected in the optimization stage. We have shown some heat maps from three iterations during the training period in Fig.10. From the figure, it is observed that using the lip loss, the lip-class probability in the background region is reduced drastically and at iteration 1200, most of the background pixels have been correctly classified.

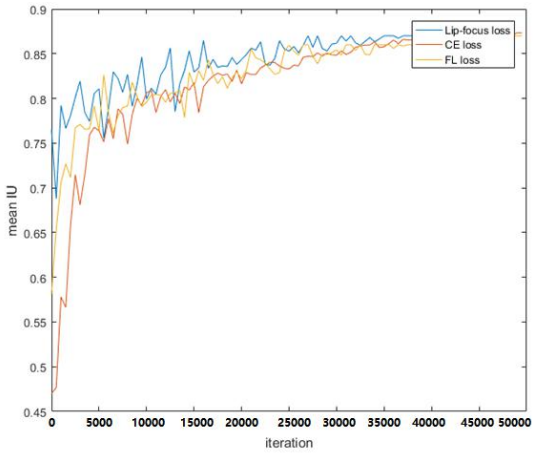

Fig.9 Convergence curves obtained by three loss functions.

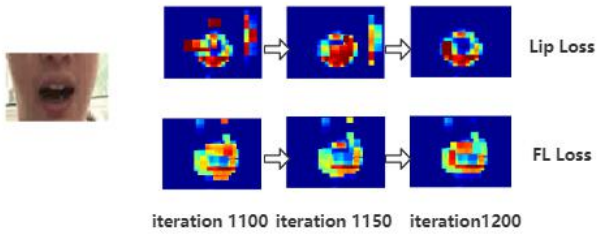

Fig.10 Heat map comparison under two loss function

C. *Effectiveness of the Fuzzy Module*

There are four fuzzy learning modules in our proposed networks to learn the intrinsic relationship between each feature map and the lip segmentation result. It should be noted that the number of membership functions would influence the fuzzy module as follows: A small number of membership functions leads to lower computational complexity while a large number of membership functions enables the fuzzy module to have more complicated fuzzy logic. Taking both the representative ability and computational complexity into consideration, a number of settings have been selected and the corresponding segmentation performance is listed in Table III ("N/A" in the table means that the corresponding fuzzy learning module is not active in the specific layer). Note that according to the network structure, the spatial size of the feature map reduces from C1 to C4. Hence, to balance the number of weights in the fuzzy module, the number of membership functions increases from C1 to C4.

From Table III, the following observations can be made. First, compared with the classical FCN structure (i.e. all the fuzzy modules are not active), the LSFCNN with fuzzy modules in one or more feature layers achieves better segmentation performance, which demonstrates the effectiveness of introducing the fuzzy learning modules. Second, when applying the fuzzy learning modules to all the feature maps, the best performance is obtained. This demonstrates that the fuzzy modules on the feature maps at different scales can learn complementary rules and taking all of them into consideration achieves the optimal performance. Third, the best segmentation performance is obtained when the number of membership functions is set as {1,2,8,8}.

To intuitively illustrate the functionality of the fuzzy modules, the output of the fuzzy modules in some channels of the feature maps is shown in Fig. 11. The black pixels in Fig.11 indicate that the output of the corresponding position is close to zero and the white pixels indicate values close to one. As shown in Fig.11, the fuzzy learning modules in the proposed network can describe the characteristics of the lip pixels at different aspects. Some modules are responsible for the upper (e.g. Module2-I) and lower lips (e.g. Module2-II), some modules describe the lip corner region (e.g. Module1-II), and some modules depict the inner mouth region (e.g. Module4-VIII). The above observations show that the fuzzy modules have learned some useful rules to differentiate the lip region against the non-lip region, and demonstrate the effectiveness of integrating the fuzzy module.

D. *Comparisons with State-of-the-Art Approaches*

To comprehensively evaluate LSFCNN, five state-of-the-art lip segmentation approaches, i.e. MS-FCM [21], MRF [16], FCN [27], LSN [25] and Mask RCNN [36] were used for comparison. Besides the above metrics, four widely used metrics in classification, i.e. sensitivity, specificity, F1-score and AUC of ROC analysis [37] were also adopted for evaluation. The lip segmentation performance using all these

TABLE III
LIP SEGMENTATION PERFORMANCE BY VARIOUS NUMBER OF MEMBERSHIP FUNCTIONS IN EACH FUZZY MODULE

Setting Index		1	2	3	4	5	6	7	8
Number of membership functions	Fuzzy module 1 (M_1)	N/A	N/A	N/A	1	1	1	1	2
	Fuzzy module 2 (M_2)	N/A	N/A	N/A	1	1	2	2	2
	Fuzzy module 3 (M_3)	N/A	N/A	8	1	2	4	8	8
	Fuzzy module 4 (M_4)	N/A	32	16	4	4	4	8	8
Segmentation performance	Accuracy	96.9%	97.9%	97.5%	98.0%	98.1%	98.2%	**98.4%**	98.3%
	mean accuracy	90.4%	90.8%	89.5%	91.4%	91.5%	92.9%	**94.5%**	**94.5%**
	mean IU	83.4%	84.1%	81.8%	85.0%	85.3%	85.8%	**87.6%**	87.4%
	frequency weighted IU	93.7%	96.1%	95.5%	96.3%	96.4%	96.5%	**96.9%**	96.7%
Time Cost	Training Time	10h	10h	10.5h	14h	14h	14h	14h	17h

TABLE IV
COMPARISON OF DIFFERENT METHODS

	MS-FCM	MRF	FCN	LSN	Mask-RCNN	**LSFCNN(ours)**
Accuracy	95.8%	96.5%	97.2%	98.1%	98.3%	**98.4%**
mean accuracy	90.2%	90.9%	92.9%	93.9%	94.2%	**94.5%**
mean IU	82.9%	83.1%	85.5%	86.1%	87.1%	**87.6%**
frequency weighted IU	91.0%	91.9%	93.2%	95.9%	96.5%	**96.9%**
Sensitivity	81.6%	82.5%	81.5%	83.6%	83.9%	**83.9%**
Specificity	94.3%	92.9%	98.5%	98.5%	98.7%	**98.9%**
F1-score	78.3%	79.0%	79.2%	80.3%	80.4%	**81.8%**
AUC of ROC	97.0%	97.1%	98.1%	98.4%	98.6%	**98.7%**

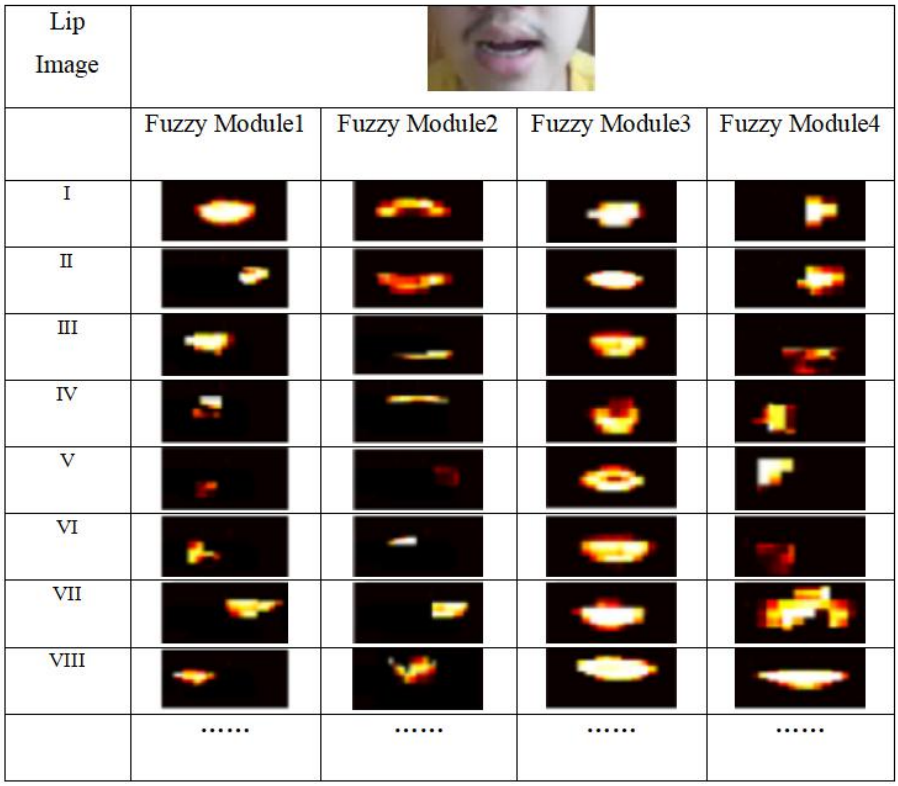

Fig. 11. Heatmaps of the fuzzy modules.

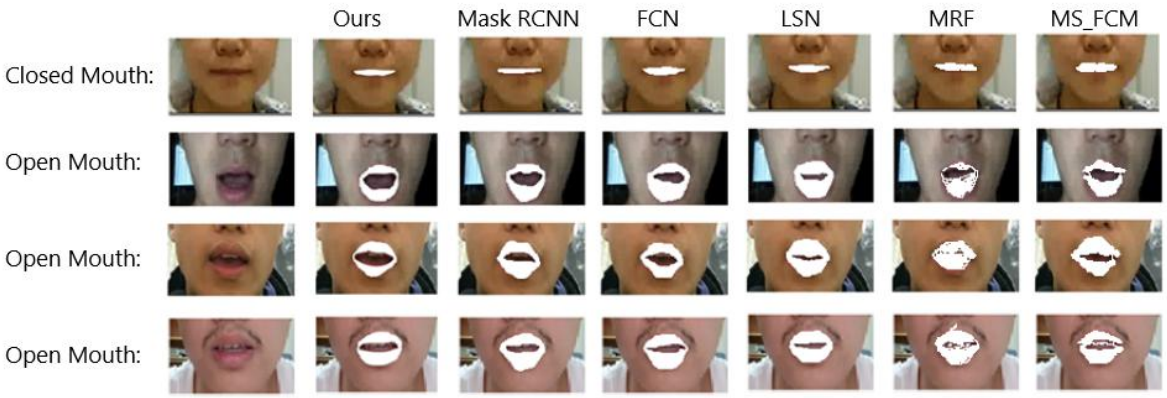

Fig.12 Lip image segmentation results using all the approaches investigated.

approaches are listed in Table IV. From the table, the following two observations can be made. Firstly, the deep learning based methods, i.e. FCN [26], LSN [25], Mask RCNN [36] and the LSFCNN achieve better segmentation performance when compared with the traditional unsupervised [16] or semi-unsupervised [18] approaches. It indicates that compared with the handcrafted rules or features, the deep neural network can extract more complex and discriminative features to differentiate lip pixels from the background pixels. Second, LSFCNN also outperformed the other three CNN-based approaches. By appropriately integrating fuzzy logic with convolutional feature maps, the intrinsic characteristics of various kinds of lip pixels and their complex relationships have been learned. Hence, even when dealing with the challenging open-mouth scenario, the proposed approach can achieve satisfactory results.

The segmentation results for four example lip images obtained by all the approaches investigated are given in Fig. 12. From the figure, it is observed that: i) for the close-mouth-typed image, all the approaches can achieve acceptable segmentation results. It is mainly because in this image, the lip pixels can be easily differentiated based on the color and spatial information; ii) for the open-mouth-typed images, owing to the complicated distribution of the lip pixels in both the color and spatial domain, the traditional approaches [21,16] may not obtain satisfactory results. With the aid of the additional supervised information (manually annotated lip images), the deep learning based approaches [25,26,36] (including ours) usually performed better; iii) Among all the deep learning based approaches, the proposed approach can achieve the best inner mouth segmentation results for all the open-mouth-typed images (esp. for the last one). It is mainly because the fuzzy modules in the proposed network can better handle the uncertainties in human annotations and the segmentation results can be more robust.

V. Conclusion

In this paper, a new novel architecture of fuzzy convolutional neural network was proposed. An end-to-end, supervised learning of convolutional neural network structure combined with fuzzy representations was proposed which can learn high level semantics. The fuzzy neural network helps the convolutional neural network to pay more attention to the lip region. With the new architecture, we achieve state-of-the-art performance with a 98.4% pixel-wise accuracy.

References

[1] I. Matthews, T.F. Cootes, J. A. Bangham J A, et al. “Extraction of visual features for lipreading.” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 2, pp. 198-213, Feb 2002.

[2] S.L. Wang, A. W. C. Liew, W. H. Lau, et al. “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. Cetingul, Y. Yemez, E. Erzin and A. M. Teklap, “Discriminative Analysis of Lip Motion Features for Speaker Identification and Speech-Reading,” *IEEE Transactions on Image Processing*, vol. 15, issue 10, pp. 2879-2891, Oct. 2006.

[4] X. Liu and Y. M. Cheung, “Learning Multi-Boosted HMMs for Lip-Password Based Speaker Verification,” *IEEE Trans. on Information Forensics and Security*, vol.9, no.2, pp.233-246, Feb.2014.

[5] C. H. Chan, B. Goswami, J. Kittler, and W. Christmas, “Local Ordinal Contrast Pattern Histograms for Spatiotemporal, LipBased Speaker Authentication,” *IEEE Transactions on Information Forensics and Security*, vol. 7-2, pp. 602-612, 2012.

[6] G. N. Kodandaramaiah, M. B. Manjunatha, S. A. K. Jilani, M. N. Giriprasad, R. B. Kulkarni and M. Mukunda Rao, "Use of lip synchronization by hearing impaired using digital image processing for enhanced perception of speech," *Proceedings of 2009 2nd International Conference on Computer, Control and Communication*, Karachi, 2009, pp. 1-7.

[7] R. Rohani, S. Alizadeh, F. Sobhanmanesh and R. Boostani, "Lip segmentation in color images," *Proceedings of 2008 International Conference on Innovations in Information Technology*, Al Ain, 2008, pp. 747-750.

[8] N. Eveno, A. Caplier, and P. Y. Coulon, “New color transformation for lips segmentation,” *Proceedings of IEEE 4th Workshop on Multimedia Signal Processing*, Cannes, France, pp.3-8, Oct. 2001.

[9] M. Shemshaki and R. Amjadifard, “Lip Segmentation

Using Geometrical Model of Color Distribution", *Proceedings of 2011 Iranian Machine Vision and Image Processing*, Tehran, Iran, pp. 1-5, Nov. 2011.

[10] Y. P. Guan, "Automatic extraction of lips based on multi-scale wavelet edge detection", *IET Computer Vision*, vol.2, issue 1, pp.23-33, March 2008.

[11] S. R. Banimahd and H. Ebrahimnezhad, "Lip Segmentation Using Level Set Method: Fusing Landmark Edge Distance and Image Information", *Proceedings of 20th International Conference on Pattern Recognition (ICPR'10)*, Istanbul, Turkey, pp. 2432-2435, Aug. 2010.

[12] A.W.C. Liew, S.H. Leung and W.H. Lau, "Lip Contour Extraction from Color Images Using a Deformable Model", *Pattern Recognition*, vol. 35, no. 12, pp. 2949-2962, 2002.

[13] S. W. Chin, K. P. Seng and L. M. Ang, "Lips Contour Detection and Tracking Using Watershed Region-Based Active Contour Model and Modified H ∞ ", *IEEE Trans. on Circuits and Systems for Video Technology*, vol.22, issue 6, pp. 869-874, 2012.

[14] C. Santiago, J. C. Nascimento and J.S. Marques, "2D Segmentation Using a Robust Active Shape Model with the EM Algorithm", *IEEE Trans. on Image Processing*, vol.24, issue 8, pp. 2592-2601, 2015.

[15] T. F. Cootes, G. J. Edwards and C. J. Taylor, "Active appearance models", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol.23, issue 6, pp.681-685, June 2001.

[16] Y. M. Cheung, M. Li, X. C. Cao, and X. G. You, "Lip Segmentation under MAP-MRF Framework with Automatic Selection of Local Observation Scale and Number of Segments", *IEEE Trans. on Image Processing*, vol. 23, issue 8, pp. 3397-3411, 2014.

[17] M. Lievin and F. Luthon, "Lip features automatic extraction," *Proceedings of 1998 International Conference on Image Processing. ICIP98 (Cat. No.98CB36269)*, Chicago, IL, USA, 1998, pp. 168-172 vol.3.

[18] X. Zhang, R. M. Mersereau, "Lip feature extraction towards an automatic speechreading system," *Proceedings of IEEE International Conference on Image Processing*, September 2000, Vancouver, BC, Canada, vol. 3, pp.226–229.

[19] A.W.C. Liew, S. H. Leung, and W. H. Lau, "Segmentation of color lip images by spatial fuzzy clustering," *IEEE Trans. Fuzzy Syst.*, vol. 11, no. 4, pp. 542–549, Aug. 2003.

[20] 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, Jan. 2004.

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

[22] J. Fu, S. Wang and X. Lin, "Robust Lip Region Segmentation Based on Competitive FCM Clustering," *Proceedings of 2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, Gold Coast, QLD, pp. 1-8, 2016.

[23] A. Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." In *Advances in neural information processing systems*, pp. 1097-1105. 2012.

[24] I. Goodfellow, et al. "Deep learning," Vol. 1. *Cambridge: MIT press*, 2016.

[25] C. Szegedy et al., "Going deeper with convolutions," *Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, pp. 1-9, 2015.

[26] Z. Ju, X. Lin, F. Li and S. Wang, "Lip Segmentation with Muti-scale Features Based on Fully Convolution Network," *Proceedings of 2018 IEEE Third International Conference on Data Science in Cyberspace (DSC)*, Guangzhou, pp. 365-370, 2018.

[27] E. Shelhamer, J. Long and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 640-651, April 2017.

[28] C. F. Juang, Y. W. Tsao, "A Self-Evolving Interval Type-2 Fuzzy Neural Network With Online Structure and Parameter Learning", *IEEE Transactions on Fuzzy Systems*, vol. 16, issue 6, pp. 1411-1424, 2008.

[29] M. Yeganejou and S. Dick, "Classification via Deep Fuzzy c-Means Clustering," *Proceedings of 2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, Rio de Janeiro, pp. 1-6, 2018.

[30] 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.

[31] 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.

[32] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, pp. 770-778, 2016.

[33] T. Lin, P. Goyal, R. Girshick, K. He and P. Dollár, "Focal Loss for Dense Object Detection," *Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, pp. 2999-3007, 2017.

[34] L. Bottou, "Large-scale machine learning with stochastic gradient descent." *Proceedings of COMPSTAT'2010*, Physica-Verlag HD, pp. 177-186, 2010.

[35] S. Santurkar, D. Tsipras, A. Ilyas, and A. Madry. "How does batch normalization help optimization?" *Advances in Neural Information Processing Systems*, pp. 2483-2493, 2018.

[36] K. He, G. Gkioxari, P. Dollár and R. Girshick, "Mask R-CNN", *Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, pp. 2980-2988, 2017.

[37] P. A. Flach, "The geometry of ROC space: understanding machine learning metrics through ROC isometrics", *Proceedings of the 20th international conference on machine learning*. pp. 194-201. 2003.