Skin Lesion Segmentation with Codec Structure Based Upper and Lower Layer Feature Fusion Mechanism

Cheng Yang[1,2*], GuanMing Lu[1]
[1]College of Telecommunications and information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China
[e-mail: 2019010214@njupt.edu.cn]
[e-mail: lugm@njupt.edu.cn]
[2]Changzhou College of Information Technology, Changzhou 213164, China
*Correspondence: 2019010214@njupt.edu.cn

Received June 12, 2021; revised October 12, 2021; accepted November 25, 2021; published January 31, 2022

Abstract

The U-Net architecture-based segmentation models attained remarkable performance in numerous medical image segmentation missions like skin lesion segmentation. Nevertheless, the resolution gradually decreases and the loss of spatial information increases with deeper network. The fusion of adjacent layers is not enough to make up for the lost spatial information, thus resulting in errors of segmentation boundary so as to decline the accuracy of segmentation. To tackle the issue, we propose a new deep learning-based segmentation model. In the decoding stage, the feature channels of each decoding unit are concatenated with all the feature channels of the upper coding unit. Which is done in order to ensure the segmentation effect by integrating spatial and semantic information, and promotes the robustness and generalization of our model by combining the atrous spatial pyramid pooling (ASPP) module and channel attention module (CAM). Extensive experiments on ISIC2016 and ISIC2017 common datasets proved that our model implements well and outperforms compared segmentation models for skin lesion segmentation.

Keywords: semantic segmentation, skin lesion segmentation, deep learning, convolutional neural network (CNN), atrous spatial pyramid pooling.

1. Introduction

Skin cancer is one of the most common malignant tumors in humans [1], and non-melanoma skin cancers are the most common malignant skin lesions. According to GLOBOCAN 2020 [2], melanoma accounts for about 20% of all skin cancer cases, while 47.2% of skin cancer deaths are due to melanoma, thus early detection of melanoma can save lives. The current methods used to identify melanoma skin lesion mainly include dermatoscopy and visual observation. Dermatoscopy, which is often used in skin imaging, has the advantages of high safety, non-invasiveness and effectiveness. However, relying only on these methods to diagnose and accurately localize melanoma skin lesions is not only labor-intensive and time-consuming, but also leads to the subjective results. Therefore, there is an urgent need to develop an intelligent end-to-end automatic segmentation method for skin lesions, which can generate accurate segmentation of skin lesions, so as to improve the diagnosis ability of doctors. Nonetheless, it is an extremely challenging task due to the complex problems of blurred boundaries, diverse appearances of lesions, low contrast between lesions and surrounding normal skin, and in some images, lesions are covered by hairs, borders, blood vessels, bubbles, *etc*.

Early on, a very popular method used to segment lesion area based on a threshold [3]. These techniques perform well in low-level segmentation tasks, but result in over-segmentation for images with low contrast, diverse colors and unclear boundary. Another important segmentation method is based on area [4], which utilizes the characteristic that the nearby pixels have uniform colors. Booming of artificial intelligence (AI) and remarkable performance of deep learning methods based on image processing tasks, researchers have started to apply it into segmentation of medical images, and have achieved huge success.

VGGNet [5], GoogLeNet [6,7], and ResNet [8] exhibited outstanding performance in image recognition tasks. Accordingly, to train better network models and improve the segmentation precision, more and more researchers prefer to use these classic networks as the backbone network, and they even use the pre-trained weights of these models (trained on the ImageNet dataset [9]) to extract features from dermoscopy images. Since the introduction of U-Net structure [10], many U-Net structure-based segmentation models have achieved remarkable indices. For example, the winner of the International Skin Imaging Collaboration (ISIC) 2018 [11] used the U-Net network structure with ResNet-101 as the backbone for feature extraction. The U-Net segmentation framework has a classic coding and decoding structure, which favors recreating the limited ability of reconstructing precise details in segmented image, due to insufficient resolution of the advanced encoder feature map, by concatenating and fusing channels among adjacent layers. Progressively deepening network will decrease resolution of feature maps and cause inadequacy of spatial information, which will lead to an insufficient positive effect of the above integration mechanism to make up for the lost spatial information, ultimately resulting in an error on the segmentation boundary.

Low-level features have abundant details, but lack semantic information. In comparison, high-level features have strong semantics, but suffer from severe loss of spatial information. The method proposed here concatenates the feature channels of each decoding unit to the feature channel of all upper-layer encoding units during the decoding stage, which can effectively integrate the spatial information and semantic information to produce better segmentation results. Furthermore, the atrous spatial pyramid pooling (ASPP) module and channel attention module (CAM) are appended to the network to enhance robustness and generalization of the model.

The vital contributions of the paper are in three aspects:

- It proposes a new deep learning-based segmentation model, which ensures the effectiveness of the segmentation by effectively integrating spatial and semantic information in decoding stage of the network. The remarkable performance and superiority of the proposed model compared to other segmentation models are demonstrated through extensive experiments on the ISIC2016 and ISIC2017 common datasets.
- By adding an ASPP module to the network, the robustness of the segmentation model for multi-scale scenarios is enhanced, and the receptive field is enlarged without increasing parameters.
- By inserting a CAM into the network, the model assigns different weights to each feature channel to make the training more focused without extra computation and storage costs.

2. Related Works

In the past few decades, many classic algorithms [12-15] have been used in skin lesion segmentation, but none of these methods can capture advanced semantic information due to their dependence on artificial features. In recent years, deep-learning has promoted the blossom of semantic segmentation methods based on it. A fully convolutional network (FCN) [17] is the pioneering work in this field which focuses on building a FCN, by inputting an image of any size, the network will output segmentation result at the same scale after effective learning and inference. Such CNN-based training model, characterized by the pattern of end-to-end and pixel-to-pixel, outperforms all previous semantic segmentation approaches, which simultaneously revealed a new direction for subsequent improvement and development of semantic segmentation algorithms. U-Net defines a contraction path to obtain the global information, and also defines a symmetrical expansion path to achieve precise positioning, from which the coding and decoding structure is established, providing end-to-end training with a small number of images with a fast-processing speed. By combining both the low-resolution and high-resolution information, U-Net is applicable to the segmentation of medical images, and has become the benchmark model in many medical imaging semantic segmentation tasks. By employing classic CNNs as the main network such as VGGNet, ResNet and Xception [16], the DeepLab series models [17-20] share the pre-trained parameter weights of the model and shorten the training time. By introducing atrous pooling, the problems of low resolution and feature extraction under multi-scale is solved. The ASPP module is designed to achieve the best performance of atrous pooling, which can improve the robustness of the network with multi-scale and multi-class segmentation.

The deep learning methods have achieved great success in segmentation of skin lesions. In [21], the appearance and context information of the image is combined in an automatic context scheme to check the boundary of the lesion. In [22], a multi-stage FCN architecture is designed, and the context information is used to repeatedly check the lesion area. [23] used the Jaccard distance loss to train the traditional FCN, so as to segment the skin lesion. In report [24], during the U-Net coding and decoding process for skin lesion detection, the dense convolution blocks are used to replace the traditional convolutional layers.

3. Proposed Method

3.1 Network architecture

The overall architecture proposed, shown in **Fig. 1**, adopts the coding and decoding structure, and it is called the Upper and Lower Layers Feature Fusing Network (ULFFN). The entire network consists of four parts: input, coding, decoding and output. The input section is completed with one convolutional operation, by which the original 3-channel color image is nonlinearly activated into 64-channel feature maps and imported into the encoder. Downsampling and feature extraction are performed on the input image in the coding, then upsampling above the feature maps is performed layer-by-layer during decoding, and eventually segmented images are generated. The model designs four downsampling and four upsampling modules as the components of the encoder and decoder, respectively. The feature map output from the decoder enters the CAM, in this way, different levels of channels can be played to focus the segmentation more on the target. Finally, the feature map with 2 channels is produced through convolution operation, and the probability of each pixel at the background and foreground is generated independently by using the SoftMax activation function to generate the binary classification score map.

Each code unit consists of a convolution block during coding, and the convolution block is composed of the deep convolution with residual structure and the downsampling operation. Pooling is widely used to downsample, but a lot of information tends to be lost during pooling. Furthermore, it cannot achieve reverse gradient updating, so pooling is not learnable. The ULFFN explores full convolution with extended convolution stride to implement downsampling in order to maximize the integrity of the image information. In a CNN, the size of the original image area mapped by the pixels on the feature map output by each layer is called the receptive field. In general, a bigger receptive field provides better results than a small one, but a bigger stride may lower the feature resolution, which will result in less information retained in the feature map. In order to maintain the same feature resolution as well as to further expand the receptive field, we embed the ASPP module at the bottom of the ULFFN to achieve richer semantic features, increasing pixel classification precision and improving segmentation performance.

In the process of decoding, the feature channel of each decode unit is connected and integrated into the feature channel of all upper-layer code units, and then upsampling is conducted, which ensures that spatial and semantic information are fully merged. Previous studies generally implemented upsampling by means of deconvolution or un-pooling, whereas we adopt the sub-pixel convolution method to acquire multiple upsampling on consideration of outstanding performances in the reconstruction of super-resolution images, and it also performed well in the training and testing of the ULFFN.

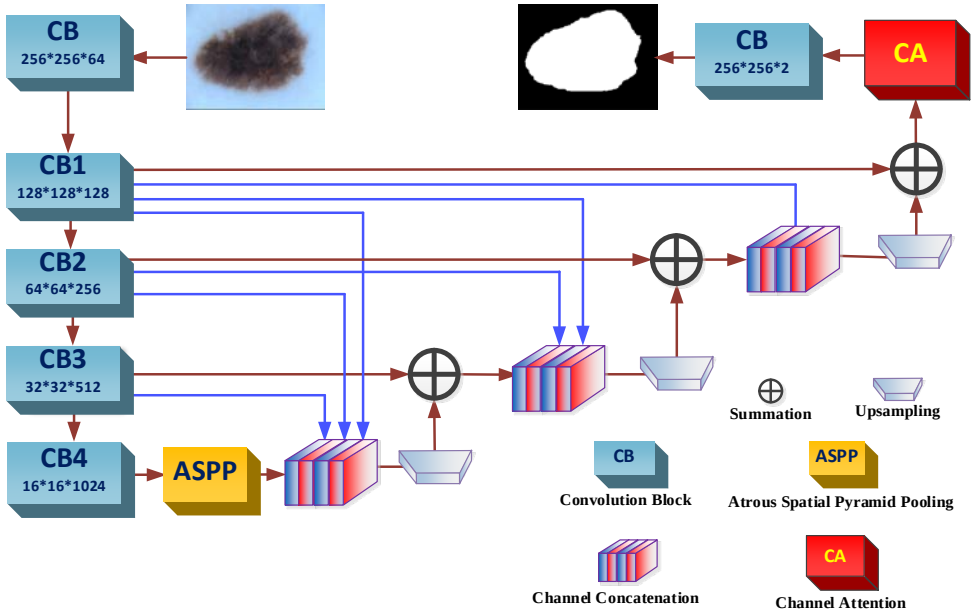


Fig. 1. The network architecture of the ULFFN.

3.2 Convolution block

The number of neural network layers plays a significant role in model performance and numerous experiments have demonstrated that deeper networks have more superior performance, as layer stacking enable the network to extract the layer features more effectively. However, if we simply superimpose layers straightway, it will cause gradient vanishing and non-convergent training processes, or the training process may converge to the local optimal minimum rather than the global optimal minimum. Moreover, increasing the number of layers may trigger a sharp convergence of the network, which will reduce the generalization ability of the model. According to a previous work [8], the deep residual structure addressed issues of gradient vanishing and sharp convergence through a skip connection, which caused gradient to directly return to the initial input layer so as to simplify the process of gradient updating by back-propagation. By which, the parameters can be kept unchanged, and the residual structure can reduce the computational complexity of the model to some extent. All four internal convolution blocks of the ULFFN use deep residual structure illustrated in **Fig. 2**. The input of convolution block fuses the output (result after two convolution operations) via a summation operation, which is achieved by skipping connection, and effectively avoids the degradation of the deep neural network. After fusion, the output performs downsampling through convolution by setting the stride to 2 to obtain a larger receptive field and extract more abundant semantic information. Each Convolution (Conv) follows the Rectified Linear Unit (ReLU) [25] as a nonlinear activation function and Batch Normalization (BN) [26], which not only enhances the learning capability of the model, but also keeps a certain gradient, thus preventing gradient vanishing or explosion. The VGGNet successfully built a deep CNN by

repeatedly stacking convolutional kernels of 3*3, and since then, many other networks have set the size of kernels as 3*3 and achieved good results, which has been shown to be an efficient and resource-saving convolution method. The receptive field with three 3*3 layers is the same as one 7*7 layer and the former only require half of the parameters of the latter. Meanwhile, the former consists of three nonlinear operations, while the latter only requires one, which means the former has stronger feature learning capacity. In ULFFN, the number of output channels of the code unit gradually increases from 128 in the first convolution block to 1,024 in the fourth convolution block, and the number of channels of one code unit is twice as many as the previous one, which exhibits a progressive increase trend from shallower to deeper layers of the network.

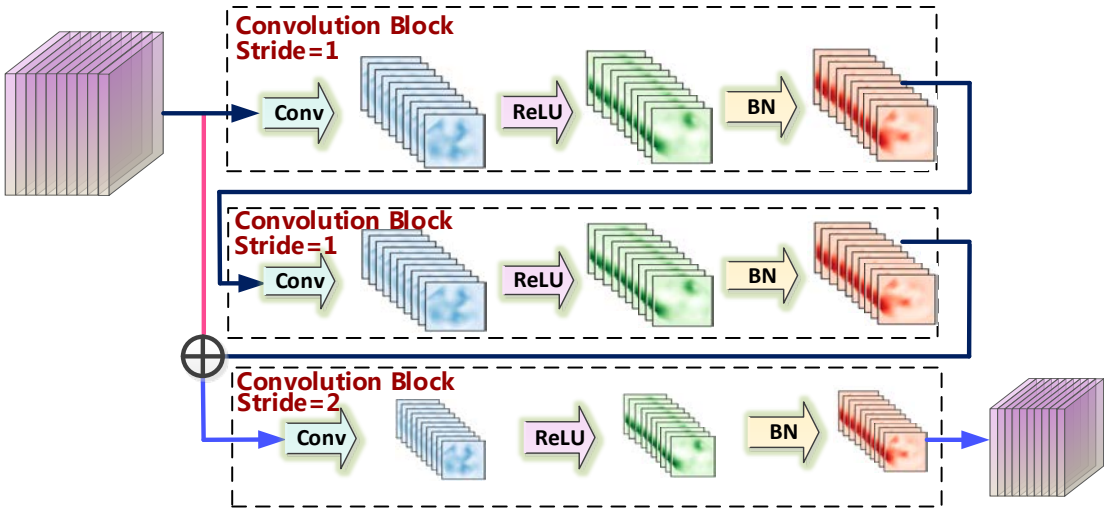


Fig. 2. Implementation details of the convolution block.

3.3 ASPP module

To image segmentation, large stride-convolution or pooling is used to lower the resolution and simultaneously increase the receptive field after that the original image is imported into the CNN with the initial ordinary convolution. As the output of the image segmentation preference is the pixel level, the downsampled image should be upsampled to the original image size for prediction, which brings about information missing due to decreased and increased resolution, *i.e.*, frequent downsampling operations, such as convolution with stride > 1 or pooling in the deep CNN will result in significantly lower spatial resolution of the feature map, which leads to damage image information. In order to overcome the obstacle and effectively generate denser feature mapping, downsampling operation is removed from the last code unit in the ULFFN, and feature mapping with a high sampling rate is achieved by implementing upsampling in the convolutional kernel of subsequent convolutional layers. Upsampling kernel involves inserting holes into the non-zero convolutional kernel and such convolution with

holes is called atrous convolution. Equations (1) and (2) describe how to calculate the receptive field and corresponding convolution kernel in size, respectively.

$$RF_{n+1} = RF_n + (k-1) \times S \tag{1}$$

$$k_{new} = k_{ori} + (k_{ori} - 1) \times (dilation\_rate - 1) \tag{2}$$

where RF denotes the size of the receptive field, k is convolutional kernel size, S represents the convolutional stride, k_{new} stands for the size of the convolutional kernel after atrous convolution, k_{ori} is original convolutional kernel size, and $dilation\_rate$ is the atrous rate. According to equation (1), under the traditional convolution operation with stride = 1, three layers of 3*3 convolution is combined to achieve a receptive field of 7*7, and the size of the receptive field is linear to layers. According to equation (2), exponentially increasing of convolutional kernel makes grow in synch with receptive field of atrous convolution.

The R-CNN [27] uses the ASPP module to resample the convolution features extracted from a single scale, which can accurately and effectively classify the region of any scale. For the given input, parallel sampling is implemented based on atrous convolution with different sampling rates, which equally captures the context information of the image with multiple size ratio. The ASPP module has improved the robustness of the network in multi-scale and multi-category segmentation, as a result of the use of different sampling ratio and receptive fields to extract the input features, which leads to the acquisition of the interesting target and context information under the conditions of various scales. The working mechanism of the ASPP module [18] is as shown in **Fig. 3**.

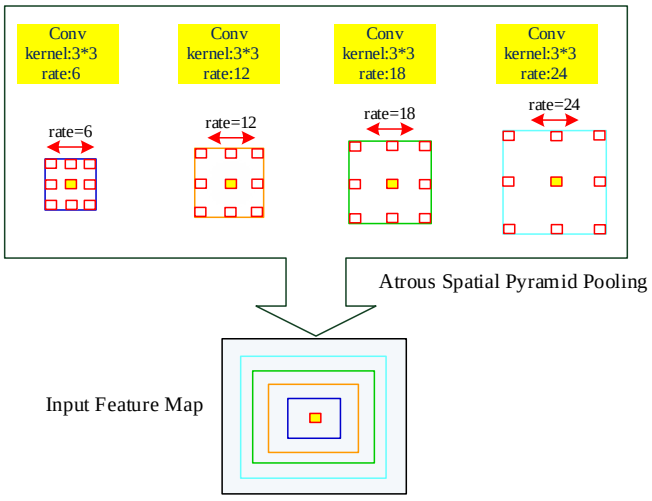


Fig. 3. The working mechanism of the ASPP module

We use multiple parallel atrous convolutions with various sampling rates to implement ASPP, further process the extracted features in separate branches according to each sampling rate, and integrate the results to generate the final result. The specific structure can be observed in **Fig. 4**.

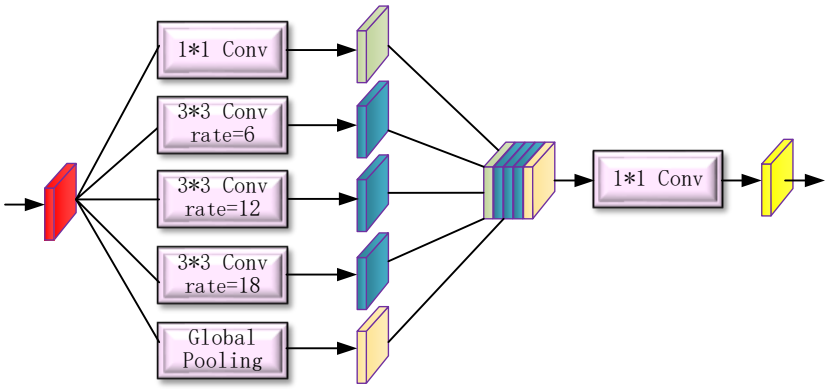


Fig. 4. The structural design of the ASPP module

3.4 Channel attention module (CAM)

The attention mechanism focuses on the target areas, while ignoring other unimportant regions and involves two steps. The chief operation is to determine which part of the input deserves more attention and then to extract features from the key part to obtain important information. The importance is determined according to the application scenario, and the neural network with the attention mechanism can conduct autonomous learning better. Analogous to the selective visual attention mechanism of mankind, attention mechanism is more interested in information related to tasks, which contributes the model by redistributing the weight of each channel and extracting significant features to improve inference capacity of the algorithm, while the consumption of calculation time and memory is hardly increased. The attention mechanism has been broadly applied in many tasks [28-30], especially in the field of computer vision. In a reported study [31], self-attention is exploited to create a better image generator, and another study [32] mainly explores the effectiveness of the non-local operation of a video and image on the time-space dimension.

The attention mechanism in computer vision are mainly categorized into three domains: spatial, channel and mixture. The Spatial Transformer Network (STN) [33] is a representative spatial attention model and essentially used to locate the target and make some transformation or obtain the weight. The Squeeze-and-Excitation Network (SENet) [34], the winner of the 2017 ImageNet Large Scale Visual Recognition Challenge (ILSVRC), is a classic channel attention model, which makes different channels of different weights, thereby preferentially allocating resources to attention channels by modeling the importance of each feature channel and then enhancing or suppressing different channels for different tasks. The Concerns-Based

Adoption Model (CBAM) [35] effectively integrates the attention of spatiality and channel and then establish mixed attention model.

Considering the outstanding performance of the SENet, the ULFFN adopts CAM to improve segmentation performance, but unlike the SENet, which mainly uses global average pooling (N_{GAP}) as squeeze operation, the ULFFN introduces global variance pooling (N_{GVP}) method, focusing on the global information, as well as the edge information, thereby the model can learn more comprehensive information. The original channel data is separately processed by N_{GAP} and N_{GVP} , and then the two types of output are fused by concatenation with one convolution. After channel activation and sigmoid activation, the new weights are calculated and the dot product between the original channel matrix and weight matrix is performed to obtain the redistribution of resources among channels, which makes model learning more focused. To guarantee the stable gradient updating and the convergence of the loss function, the residual branch is added in the CAM. The specific implementation details are shown in **Fig. 5**.

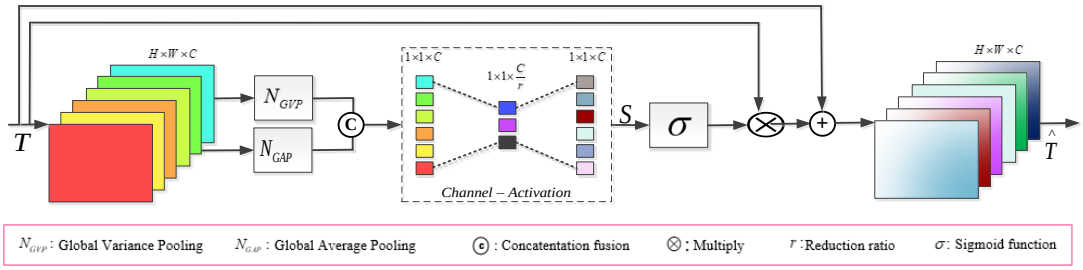


Fig. 5. CAM of the ULFFN

4. Experiment and results

4.1 Skin lesion dataset

The ISIC is an international organization dedicated to the detection of skin cancer and the ISIC Challenge focuses on analysis of skin lesions and the detection of skin cancer. This challenge includes three major tasks of lesion segmentation, detection of lesion properties and classification of skin disease. The ISIC provides the dermoscopy images and corresponding references for training of neural networks. ISIC2016 dataset [36] owns 900 images for training and 350 images for testing, respectively. While ISIC2017 dataset [37] possesses 2000 images for training and 150 images for validating, respectively.

We use ISIC2016 dataset as the training dataset for the ULFFN, which provides 900 images in JPEG format for the training of lesion segmentation tasks, including the dermoscopy images obtained from different devices in various advanced international clinical centers.

During the training process, we randomly chose 200 images from ISIC2016 training data as verification dataset. To test robustness and generalization, we evaluated the model on ISIC2016 and ISIC2017test dataset. The ISIC2016 test set was a test dataset issued by ISIC in 2016, and includes 379 biomedical images and corresponding labels. The ISIC2017 test set was issued by ISIC in 2017 for segmentation and test of lesion images, and includes 600 biomedical images and corresponding labels.

4.2 Evaluation index

From medical perspective people mainly focus on two criteria: specificity and sensitivity. Sensitivity is essentially a recall ratio which indicates whether all true positives have been found. Specificity refers to the ratio of false positives. The segmentation of skin lesion is a classic binary classification issue, and false prediction results of a model mainly include two types: one is false reporting of negative as positive (reporting a disease when there is no disease), and the other is false reporting of positive as negative (reporting no disease when there is disease). The optimization process is to simultaneously reduce these two types of errors which is essentially a process to achieve two categories of errors trade-off, thus it is meaningless to simply reduce one type of error while ignoring the other. Under the condition of binary classification, there are four possible prediction results:1)True Positive (TP), 2)True Negative (TN), 3)False Positive (FP), 4)False Negative (FN). **Table 1** expresses the details.

Table 1. The confusion matrix for binary classification

		Actual value	
		Positive	Negative
Predicted value	Positive	TP number	FP number
	Negative	FN number	TN number

From the above confusion matrix, we can obtain the following calculation formulas of index:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{3}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{4}$$

$$Specificity = \frac{TN}{FP + TN} \tag{5}$$

$$JSI = \frac{TP}{TP + FP + FN} \tag{6}$$

$$Dice = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{7}$$

Accuracy refers to the prediction accuracy and stands for the proportion of TP and TN samples to the total samples, a higher value indicates higher prediction accuracy, and is defined by equation (3). Sensitivity, i.e., the true positive rate, reflects the degree of false report, and it is defined by equation (4). Specificity refers to the proportion of accurately predicted samples in negative samples, reflecting the degree of false report, and is defined by equation (5). The JSI stands for the Jaccard similarity index and applies to compare the similarity between sample sets, *i.e.*, the Intersection over Union (IoU). The higher the JSI is, the more similar the samples are to each other, and it is defined by equation (6). The Sørensen–Dice coefficient which has similar functions as the JSI, is also used to measure the similarity between samples, its value is within the range of 0-1, and it is defined by equation (7). All the above four indices are introduced to evaluate the performance of the ULFFN.

4.3 Loss function

For most semantic segmentation scenarios, leveraging cross entropy to calculate loss for gradient update is conducive to a steady training of the model and equation (8) is its mathematical expression. For binary classification tasks, a special case of cross-entropy, namely binary cross entropy, is used, and its mathematical expression is presented in equation (9). However, in scenarios of skin lesion segmentation, the categories of segmentation only include foreground and background. Moreover, the number of pixels of the foreground is significantly lower than that of the background, i.e., the number of y=0 is remarkably higher than the number of y=1in equation (8) and dominates the loss function, which will result in a prediction that is heavily biased to the background and generate poor segmentation results. Some images that reflect the above peculiarity are shown in Fig. 6. Focal Loss [38] is the promotion to cross-entropy loss function and mainly addresses the unbalanced numbers of difficult and easy samples, which can also be interpreted as the mining of difficult samples It was initially designed to solve the severely unbalanced numbers of positive and negative samples in target detection. Parameterized cross entropy can regulate unbalanced samples (see

equation (10) for details). Most candidate targets in target detection are easy samples with little loss, which account for the majority of targets and control the loss value. Therefore, it is considered that easy samples have little effect on improving the model performance and the model should focus on the difficult samples. To address this issue, Focal Loss has evolved to equation (11). When parameter p→0, the regulatory factor (1-p) is closer to 1, and the loss is not affected; when p→1, (1-p) is closer to 0, which will reduce the contribution of easy samples to the total loss. When r=0, Focal Loss is the traditional cross entropy, but when r increases, the regulatory coefficient will increase accordingly. When r is a fixed value, such as p=2, the loss caused by easy samples is 100 times smaller than that caused by the standard cross entropy, and when p=0.968, the loss of easy samples is 1,000 times smaller than that of the standard cross entropy. However, for difficult samples (p<0.5), the difference is 4 times, so the weight of difficult samples is significantly increased, which increases the importance of misclassifications. The Dice coefficient is a judgment indicator of segmentation effect and its formula is equivalent to IoU between inference area and original area, which tackled the issue of extreme imbalance of positive and negative samples via ignoring vast background pixels when calculating IoU (see equation (12) for details). Focal Loss and Dice Loss have better effects for category imbalance issues, but for balanced samples, they increase the complexity of network training and have no advantages in performance, while cross entropy Loss function is just the opposite. During the training of our model, the original image is uniformly adjusted to the size of 512*512 by a center cropping operation before being input into the network, consequently the proportion of pixels of the foreground and background of the original image are not significantly different from each other, thus ULFFN adopts cross entropy as loss function (See equation 13 for specific calculation formula, where x is the output vector of the network and class is the label).

$$L = -\sum_{i=1}^{M} y_i \log p_i \tag{8}$$

$$CE(p, y) = \begin{cases} -\log p & y = 1 \\ -\log(1-p) & y = 0 \end{cases} \tag{9}$$

$$CE(p, y) = \begin{cases} -\alpha \log p & y = 1 \\ -(1-\alpha)\log(1-p) & y = 0 \end{cases} \tag{10}$$

$$L_{fl} = \begin{cases} -\alpha(1-p)^{\gamma} \log p & y = 1 \\ -(1-\alpha)p^{\gamma} \log(1-p) & y = 0 \end{cases} \tag{11}$$

$$L_{Dice} = 1 - \frac{2TP}{2TP + FN + FP} \tag{12}$$

$$Loss(x, class) = -log(\frac{exp(x[class])}{\sum_j exp(x[j])}) = -x[class] + log(\sum_j exp(x[j])) \tag{13}$$

Fig. 6. Skin lesion image samples (from the ISIC2016 training dataset)

4.4 Environment setting

We do not use pre-trained networks such as ResNet as backbone of our model and the training is started from scratch. Adam [32] is exploited to optimize network. The original learning rate is 1e-4 and reduced 10% every 20 epochs. Our model is trained for 100 epochs and batch-size is set to 8. Two NVIDIA GeForce RTX 2080 Ti GPU parallelly trained our model. In order to improve the model performance and reduce computation time, the images input into the network were pre-cut into the size of 512*512.

4.5 Results and analysis

To prove superiority of the proposed model, we conduct comparative analysis of ULFFN with classical models including FCN, SegNet [39], U-Net and DeepLabV3+ on ISIC2016 and ISIC2017 test datasets. FCN established fully CNN that input images in arbitrary size and generated counterpart output with powerful performance. SegNet adopted fully CNN architecture for pixel-wise segmentation comprising of an encoder-decoder network. U-Net consists of two parts, one contraction path is defined to obtain global information, and the other symmetric expansion path is defined to exact location, which can be used for end-to-end training with tiny images to receive relatively fast training. DeepLabV3+ integrated merits of ASPP and codec-structure to extend DeepLabv3 by appending a decoder to refine results illustrated in **Fig. 7**, ULFFN achieved the highest accuracy and lowest loss among all models of comparison.

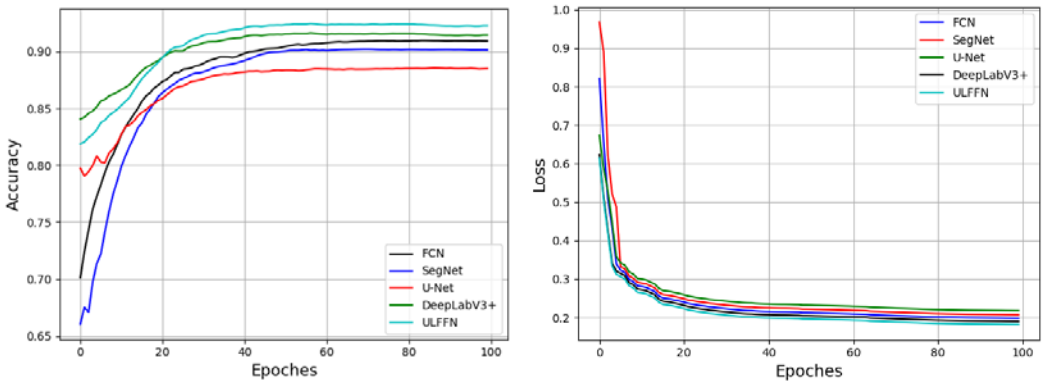


Fig. 7. The changes of all compared methods in accuracy and loss on ISIC2016 test dataset.

According to the results presented by **Table 2** and **Table 3**, ULFFN outperforms comparison of approaches on both datasets, and all metrics are higher than other comparative models. On ISIC2016 test dataset, the Accuracy, Dice coefficient and JSI values of the ULFFN reach 0.91283, 0.93044 and 0.87360, respectively, which are approximately 0.04, 0.03 and 0.04 higher than the lowest values. On ISIC2017 test dataset, the Accuracy, Dice coefficient and JSI values of ULFFN attain 0.89199, 0.9250 and 0.87085, respectively, which are about 0.04, 0.02 and 0.03 higher than the lowest values. All compared models were trained using 700 images randomly chosen from ISIC2016 training dataset, while ISIC2017 training dataset contains as many as 2,000 images of skin lesions, and its data distribution is vastly different from that of the ISIC2016 training dataset. Furthermore, there are 600 images in ISIC2017 test dataset, nearly doubling the number of ISIC2016 test dataset. Therefore, the test result on ISIC2017 test dataset was a little worse than that on ISIC2016 test dataset.

We performed ablation experiments on ISIC2016 dataset to analyze the contribution of the ASPP and CAM in improving the performance of the ULFFN. Experiments include four scenarios: 1) The basic model without integrating the ASPP or CAM (WO-AC). 2) The model integrated with the ASPP module (W-A). 3) The model integrated with the CAM (W-C). 4) The model integrated with both the ASPP and CAM (W-AC). The specific test results presented in Table 4 reveal that the model without ASPP or CAM has the lowest indices, and the values of the Accuracy, Dice coefficient and JSI are 0.88924, 0.87978 and 0.82551, respectively. On the other hand, the model with both modules has the best indices, and the values of the Accuracy, Dice and JSI are as high as 0.92876, 0.93044 and 0.87360, respectively, which indicates that the ASPP and CAM can remarkably improve model performance. The model with W-A outperforms the model with W-C in various data according to the compared test results, which indicates that the ASPP module enhances the effect of the ULFFN more than the CAM does.

The visual segmentation results on ISIC2016 and ISIC2017 test datasets are illustrated in **Fig. 8** and **Fig. 9**, respectively. The images shown comprise three parts from left to right: the

left one is the original image, the middle one is the label image, and the right one is the prediction. The size of both the original image and label image is 512*512 by performing center cropping on the original dataset, consistent with the training process. The predictions are highly close to the actual labels for different image samples, which confirm the strong robustness of the ULFFN. The effect of segmentation is slightly distinct on different test datasets, *e.g.*, the metrics are a little higher on ISIC2016 test dataset than those on ISIC2017 test dataset, which is in line with the gap between models on these two test datasets discussed above, and indicates the strong generalization ability of the ULFFN.

Table 2. Performance comparison of the ULFFN and classical semantic segmentation models on ISIC2016 test dataset.

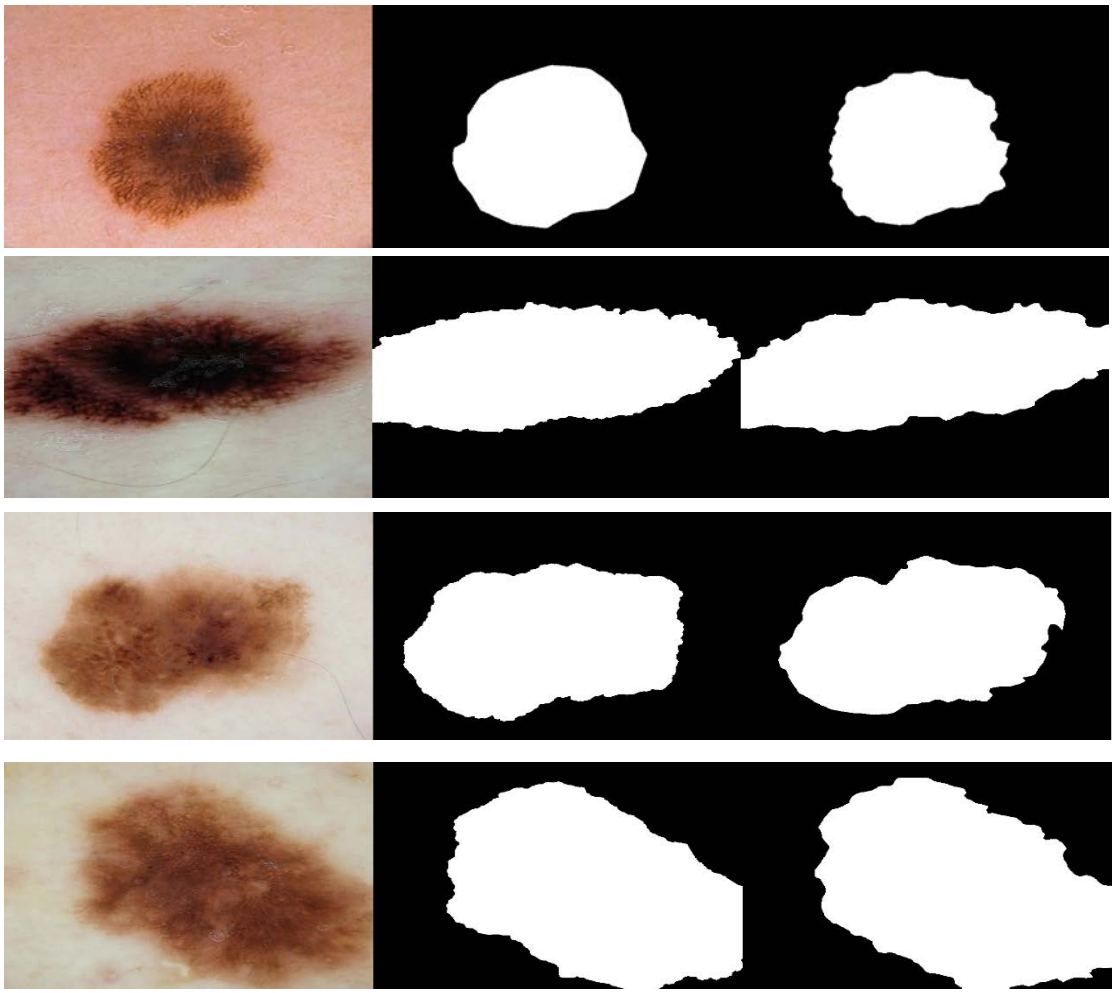
Model	Accuracy	Dice	JSI	Sensitivity	Specificity
FCN	0.90876	0.91918	0.85735	0.91333	0.83751
SegNet	0.90108	0.91374	0.84626	0.89920	0.87160
U-Net	0.88555	0.90020	0.82628	0.89208	0.80227
DeepLab v3+	0.91283	0.92317	0.86085	0.91203	0.88278
ULFFN	0.92876	0.93044	0.87360	0.92725	0.90040

Table 3. Performance comparison of ULFFN and classical semantic segmentation models on ISIC2017 test dataset.

Model	Accuracy	Dice	JSI	Sensitivity	Specificity
FCN	0.87079	0.91115	0.85062	0.91917	0.88366
SegNet	0.86081	0.90746	0.84237	0.91301	0.87257
U-Net	0.85062	0.90062	0.83202	0.91259	0.80385
DeepLab v3+	0.87579	0.91747	0.85880	0.91619	0.83879
ULFFN	0.89199	0.92502	0.87085	0.94386	0.92224

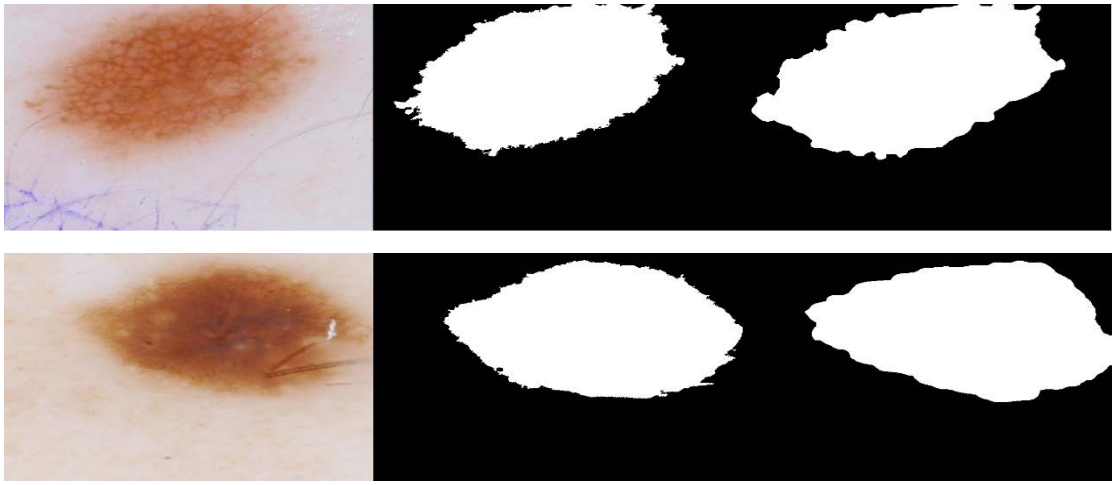
Table 4. Performance comparison of the ULFFN models without the ASPP or CAM (WO-AC), with the ASPP module (W-A), with the CAM (W-C), and with both the ASPP and CAM (W-AC).

Module	Accuracy	Dice	JSI	Sensitivity	Specificity
WO-AC	0.88924	0.87978	0.82551	0.88703	0.86894
W-A	0.90924	0.89932	0.84898	0.90782	0.88796
W-C	0.91632	0.90638	0.85647	0.91033	0.89921
W-AC	0.92876	0.93044	0.87360	0.92725	0.90040



Original image Label Prediction

Fig. 8. The result of segmentation of ULFFN on ISIC2016 test dataset.



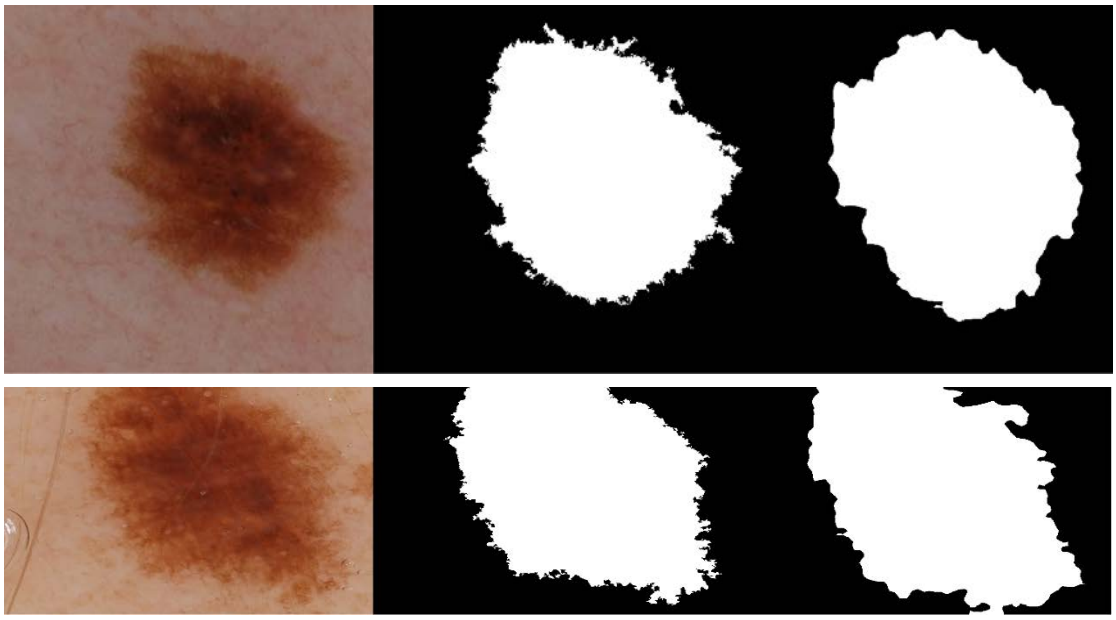


Fig. 9. The result of segmentation of ULFFN on ISIC2016 test dataset.

5. Conclusion

We proposed a new deep learning segmentation model based on the architecture of coding and decoding for skin lesion segmentation, which can effectively integrate spatial and semantic information during decoding, thus ensuring satisfactory segmentation results. The ASPP and CAM are introduced into the network to markedly enhance the robustness of the model during multi-scale segmentation. As a result, weights can be assigned to various feature channels, making the network training more focused, and improving the performance without extra computation time or memory consumption. Extensive experiments performed on public datasets ISIC2016 and ISIC2017 demonstrated that the proposed model has outstanding performances in skin lesion segmentation, and exceeds classical models of comparison and shows great robustness and generalization.

Acknowledgement

This work was supported by The Engineering Technology Research and Development Center of Jiangsu Higher Vocational and Technology Colleges, The Industrial Big Data and Intelligent Engineering Technology Research and Development Center.
This work was partly supported by the National Natural Science Foundation of China (NSFC) under Grants 72074038

References

[1] S. Pathan, K. G. Prabhu, P. C. Siddalingaswamy, "Techniques and algorithms for computer aided diagnosis of pigmented skin lesions—A review," *Biomedical Signal Processing and Control*, vol.39, pp.237-262, 2018. Article(CrossRef Link)

[2] GLOBOCAN, 2020. https://gco.iarc.fr/today/home

[3] G. Sforza, G. Castellano, S. K. Arika, R. W. LeAnder, R. J. Stanley, W. V. Stoeckerand, J. R. Hagerty, "Using adaptive thresholding and skewness correction to detect gray areas in melanoma in situ images," *IEEE Transactions on Instrumentation and Measurement*, vol.61, no.7, pp.1839-1847, 2012. Article(CrossRef Link)

[4] B. Peng, L. Zhang, D. Zhang, "Automatic image segmentation by dynamic region merging," *IEEE Transactions on image processing*, vol.20, no.12, pp.3592-3605, 2011. Article(CrossRef Link)

[5] K. Simony, A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint*, 2014. Article(CrossRef Link)

[6] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, "Going deeper with convolutions," in *Proc. of the IEEE conference on computer vision and pattern recognition*, pp.1-9, 2015. Article(CrossRef Link)

[7] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. of the IEEE conference on computer vision and pattern recognition*, pp.2818-2826, 2016. Article(CrossRef Link)

[8] K. He, X. Zhang, S. Ren, J. Sun, "Deep residual learning for image recognition," in *Proc. of the IEEE conference on computer vision and pattern recognition*, pp.770-778, 2016. Article(CrossRef Link)

[9] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol.115, no.3, pp.211-252, 2015. Article(CrossRef Link)

[10] O. Ronneberger, P. Fischer, T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. of International Conference on Medical image computing and computer-assisted intervention*, pp.234-241, 2015. Article(CrossRef Link)

[11] N. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti, H. Kittler, A. Halpern, "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic)," *arXiv preprint*, 2019. Article(CrossRef Link)

[12] R. Garnavi, M. Aldeen, M. E. Celebi, A. Bhuiyan, C. Dolianitis, G. Varigos, "Automatic segmentation of dermoscopy images using histogram thresholding on optimal color channels," *World Academy of Science, Engineering and Technology International Journal of Biomedical and Biological Engineering*, Vol. 5, No. 7, pp. 275-283, 2011. Article(CrossRef Link)

[13] A. Wong, J. Scharcanski, P. Fieguth, "Automatic skin lesion segmentation via iterative stochastic region merging," *IEEE Transactions on Information Technology in Biomedicine*, vol.15, no.6, pp.929-936, 2011. Article(CrossRef Link)

[14] Z. Liu, J. Zerubia, "Skin image illumination modeling and chromophore identification for melanoma diagnosis," *Physics in Medicine & Biology*, vol.60, no.9, pp.3415, 2015. Article(CrossRef Link)

[15] R. B. Oliveira, N. Marranghello, A. S. Pereira, J. M. R. Tavares, "A computational approach for detecting pigmented skin lesions in macroscopic images," *Expert Systems with Applications*, vol.61, pp.53-63, 2016. Article(CrossRef Link)

[16] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. of the IEEE conference on computer vision and pattern recognition*, pp.1251-1258, 2017. Article(CrossRef Link)

[17] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *arXiv preprint*, 2014. Article(CrossRef Link)

[18] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, 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, 2017. Article(CrossRef Link)
[19] L.-C. Chen, G. Papandreou, F. Schroff, H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint*, 2017. Article(CrossRef Link)
[20] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. of the European conference on computer vision (ECCV)*, pp.833-851, 2018. Article(CrossRef Link)
[21] Z. Mirikharaji, S. Izadi, J. Kawahara, G. Hamarneh, "Deep auto-context fully convolutional neural network for skin lesion segmentation," in *Proc. of 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pp.877-880, 2018. Article(CrossRef Link)
[22] L. Bi, J. Kim, E. Ahn, A. Kumar, M. Fulham, D. Feng, "Dermoscopic image segmentation via multistage fully convolutional networks," *IEEE Transactions on Biomedical Engineering*, vol.64, no.9, pp.2065-2074, 2017. Article(CrossRef Link)
[23] Y. Yuan, M. Chao, Y.-C. Lo, "Automatic skin lesion segmentation using deep fully convolutional networks with jaccard distance," *IEEE transactions on medical imaging*, vol.36, no.9, pp.1876-1886, 2017. Article(CrossRef Link)
[24] S. Vesal, N. Ravikumar, A. Maier, "SkinNet: A deep learning framework for skin lesion segmentation," in *Proc. of 2018 IEEE Nuclear Science Symposium and Medical Imaging Conference Proceedings (NSS/MIC)*, pp.1-3, 2018. Article(CrossRef Link)
[25] V. Nair, G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proc. of the 27th International Conference on International Conference on Machine Learning(ICML)*, pp.807-814, 2010. Article(CrossRef Link)
[26] S. Ioffe, "Batch renormalization: Towards reducing minibatch dependence in batch-normalized models," in *Proc. of Advances in neural information processing systems*, pp.1942-1950, 2017. Article(CrossRef Link)
[27] R. Girshick, J. Donahue, T. Darrell, J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. of the IEEE conference on computer vision and pattern recognition*, pp.580-587, 2014. Article(CrossRef Link)
[28] G. Lin, C. Shen, A. Van Den Hengel, I. Reid, "Efficient piecewise training of deep structured models for semantic segmentation," in *Proc. of the IEEE conference on computer vision and pattern recognition*, pp.3194-3203, 2016. Article(CrossRef Link)
[29] Z. Lin, M. Feng, C. N. d. Santos, M. Yu, B. Xiang, B. Zhou, Y. Bengio, "A structured self-attentive sentence embedding," *arXiv preprint*, 2017. Article(CrossRef Link)
[30] T. Shen, T. Zhou, G. Long, J. Jiang, S. Pan, C. Zhang, "Disan: Directional self-attention network for rnn/cnn-free language understanding," in *Proc. of the AAAI Conference on Artificial Intelligence*, pp. 5446-5455, 2018. Article(CrossRef Link)
[31] H. Zhang, I. Goodfellow, D. Metaxas, A. Odena, "Self-attention generative adversarial networks," in *Proc. of International conference on machine learning*, pp.7354-7363, 2019.
[32] X. Wang, R. Girshick, A. Gupta, K. He, "Non-local neural networks," in *Proc. of the IEEE conference on computer vision and pattern recognition*, pp.7794-7803, 2018. Article(CrossRef Link)
[33] M. Jaderberg, K. Simonyan, A. Zisserman, "Spatial transformer networks," *Advances in neural information processing systems*, vol.2, pp.2017-2025, 2015.Article(CrossRef Link)
[34] J. Hu, L. Shen, G. Sun, "Squeeze-and-excitation networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 8, pp. 2011-2023, 2020. Article(CrossRef Link)
[35] S. Woo, J. Park, J.-Y. Lee, I. So Kweon, "Cbam: Convolutional block attention module," in *Proc. of the European conference on computer vision (ECCV)*, pp.3-19, 2018. Article(CrossRef Link)
[36] D. Gutman, N. C. Codella, E. Celebi, B. Helba, M. Marchetti, N. Mishra, A. Halpern, "Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (ISBI) 2016, hosted by the international skin imaging collaboration (ISIC)," *arXiv preprint*, 2016. Article(CrossRef Link)

[37] N. C. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, "Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic)," in *Proc. of 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pp.168-172, 2018. Article(CrossRef Link)

[38] T. -Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, "Focal loss for dense object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318-327, 2020. Article(CrossRef Link)

[39] V. Badrinarayanan, A. Kendall, R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481-2495, 2017. Article(CrossRef Link)

Cheng Yang received the B.E. degree in computer science and technology from Jiangsu University of Technology, Changzhou, China, in 1998, and the M.E. degree in computer application technology from Jiangsu University, Zhenjiang, China, in 2007. He is currently pursuing the Ph.D. degree in Nanjing University of Posts and Telecommunications (NJUPT), Nanjing, China. He works in deep learning and computer vision such as image reconstruction, image recognition, image segmentation, and target detection.

GuanMing LU received the B.S. degree in radio engineering from Nanjing University of Posts & Telecommunications, Nanjing, China, in 1985, and the M.S. degree in communication and electronic system from Nanjing University of Posts & Telecommunications, Nanjing, China, in 1988, and the Ph.D. degree in communication and information system from Shanghai Jiaotong University, Shanghai, China, in 1999. He has been engaged in image processing, multi-media communications, digital television, an expression recognition.