ISSN 2616-5775 Vol. 7, Issue 9: 60-66, DOI: 10.25236/AJCIS.2024.070908

Medical Image Recognition Based on Multiscale Cascade Segmentation Network MCSnet

Yucheng Liu[1], Shuchang Huang[2,*], Zijun Wang[3]

[1]*College of Mathematics and Physics, Chengdu University of Technology, Yibin, China*
[2]*College of Energy, Chengdu University of Technology, Chengdu, China*
[3]*College of Materials and Chemistry & Chemical Engineering, Chengdu University of Technology, Yibin, China*
[*]*Corresponding author: huangshuchang03@163.com*

Abstract: *In this study, a new deep learning model-Multi-scale Cascade Segmentation Network (MCSnet)-is proposed for the automatic analysis of lung radiographs and pneumonia detection. MCSnet combines an encoder, an ASPP module, and a decoder to efficiently extract the multiscale semantic information and realize the accurate recognition of lung abnormalities. The experimental results on the qata_v2 dataset show that the MCSnet model has excellent performance, with an average accuracy of 92.81% and an* $mIoU$ *of 86.33%, which is a significant enhancement compared to the traditional methods. With the multi-scale segmentation technique, the method in this study is able to comprehensively capture the details of lung lesions, providing reliable support for the diagnosis of pneumonia and bringing new possibilities for clinical diagnosis.*

Keywords: *Medical Image Segmentation, Deep Learning, Multiscale Features*

1. Introduction

In today's era of big data and artificial intelligence, it is crucial to diagnose lung diseases accurately and quickly [1]. The complex diversity of pneumonia makes it challenging to diagnose, especially in lung radiograph analysis. Traditional methods suffer from recognition difficulties and inefficiency, and there is an urgent need for smarter and more adaptable image analysis methods. To address this need, this study proposes an innovative deep learning model-Multi-scale Cascade Segmentation Network (MCSnet)-that combines an encoder, an ASPP module, and a decoder, and is able to accurately extract the multi-scale semantic information to realize the accurate recognition of lung abnormalities. The experimental results show that the MCSnet model exhibits excellent performance in pneumonia detection, with an average accuracy of 92.81% and an $mIoU$ of 86.33%, which is significantly improved compared to the traditional methods. Through this research method, new possibilities can be brought for pneumonia diagnosis and provide reliable support and diagnostic basis for clinicians.

2. Data description and data preprocessing

2.1 Description of the data set

The main dataset used for the experiments in this paper is the QaTa-COV19-v2 dataset. QaTa-COV19-v2 dataset is a set of datasets consisting of lung X-ray images, created by Qatar University, University of Tampere and Hamad Medical Corporation for training and evaluating computer vision models.

The QaTa-COV19-v2 dataset contains 9258 chest images and all images are cropped to 224×224. 9258 images were selected to be used as the dataset for this study and divided the dataset into training, testing and validation sets in a ratio of 8:1:1 as shown in Table. 1. Also, each image has a corresponding mask that is used to identify the abnormal lung shadows present in the image. These masks are binary mask maps corresponding to the images and are of the same size as the original image, where the abnormal lung shadows are labeled as white, and the background is labeled as black.

The QaTa-COV19-v2 dataset contains lung radiographs of both genders at all ages, and therefore the models trained using this dataset have higher reliability and generalization in the clinic. Some of the

ISSN 2616-5775 Vol. 7, Issue 9: 60-66, DOI: 10.25236/AJCIS.2024.070908

dataset images are shown in Figure 1.

Table 1: Experimental data set information

	Image	Tags
Training Set	7406	7406
Validation Set	926	926
Test Set	926	926
Total	9258	9258

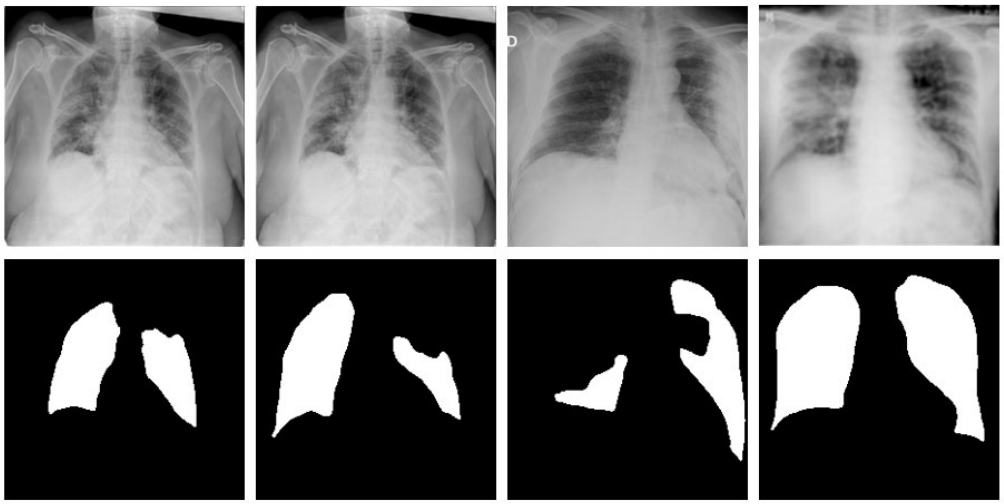

Figure 1: Partial training set images-lung X-ray proto-images and corresponding proto-masks

2.2 Data preprocessing

Data enhancement is used to generate new training samples by applying a series of random transformations to the original image [2]. This experiment performed the following types of data enhancement operations on the original images as shown in Figure 2.

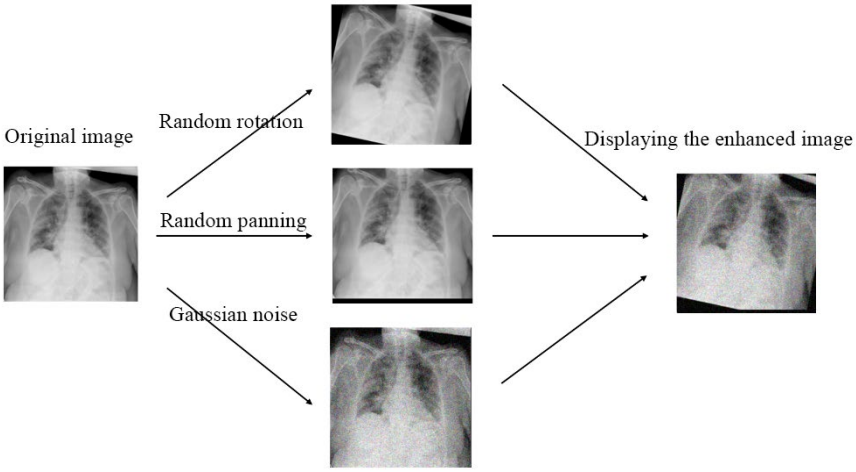

Figure 2: Image enhancement

Random Rotation: a random angle is chosen from -45° to 45° and there is a 50% probability that the image will be rotated during the training process. This can simulate the observation of the image at different angles and increase the diversity of the training data.

Random panning: with a 30% probability, a randomly selected panning distance within a specified range is applied to the image. This can simulate the observation of the image under different positions and increase the diversity of the training data.

Gaussian Noise: Generate random Gaussian distributed noise with a 20% probability of adding it to the image. This increases the complexity of the image and allows the model to learn better resistance to

ISSN 2616-5775 Vol. 7, Issue 9: 60-66, DOI: 10.25236/AJCIS.2024.070908

noise.

With these image enhancement operations, more and more diverse training samples can be generated, which helps to improve the generalization ability and performance of the model.

3. Multi-scale cascaded segmentation network

In this paper, a multi-scale cascaded segmentation network MCSnet (Multi-scale Cascaded Segmentation network) is proposed, MCSnet cascades multiple model network structures to improve the accuracy and efficiency of segmentation. As shown in Figure. 3, the MCSnet network structure is mainly composed of three parts: encoder, atrous space pyramid pooling module (ASPP) and decoder. Firstly, the input 224×224 image X is subjected to feature extraction by the Encoder network to obtain the semantic feature Y. Then, the feature Y is subjected to the ASPP module to capture the context information at different scales to obtain the multi-scale semantic feature Z. Finally, the Decoder module refines the Z features, restores the resolution, and obtains the segmentation prediction map W.

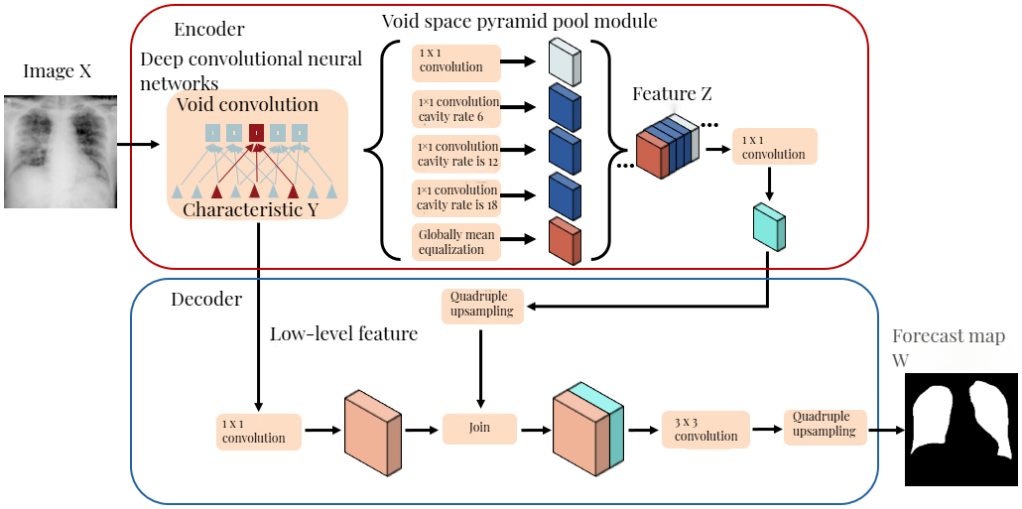

Figure 3: MCSnet network structure diagram

3.1 Encoder

The decoder-encoder architecture is an End-to-End deep learning framework [3], where the encoder and decoder play different roles in the model, with the encoder being responsible for capturing the deep features and global contextual information of the image, while the decoder is responsible for recovering the spatial details of the image and performing fine-grained segmentation. This design enables MCSnet to achieve excellent performance when processing complex images.

Xception is used as the backbone network for feature extraction. The structure of Xception network mainly consists of an input layer (Entry flow) for feature extraction and dimensionality reduction of the input feature maps, an intermediate layer (Middle flow) for extraction of more complex features, and an output layer (Exit flow) for classification of the feature maps.

3.2 Dilated space pyramid pooling

In order to make the MCSnet model better able to handle targets of different sizes, the segmentation performance of the model by introducing atrous space pyramid pooling (ASPP)is enhanced, which performs feature extraction at different void convolution steps, and fuses features of different scales together.

Dilated convolution, also known as inflationary convolution, is a widely used technique in computer vision [4], which enables the model to expand the sensory field size exponentially and linearly without increasing the computational effort by introducing the dilated rate on top of the traditional convolution. The core idea is to adjust the spacing between elements in the convolution kernel by increasing the void parameter, thus effectively expanding the receptive field. This method can not only effectively solve the problem of feature image resolution, but also help reduce the burden of model training. The formula for

ISSN 2616-5775 Vol. 7, Issue 9: 60-66, DOI: 10.25236/AJCIS.2024.070908

calculating the size of the dilated convolution kernel is as follows.

$$K = k + (k-1)(r-1) \tag{1}$$

Where k is the size of the normal convolution kernel, K is the size of the dilated convolution kernel, and r is the dilated rate.

The structure of the dilated convolution is shown in Figure 4. The entire grid in the figure represents the input of the lung X-ray image, the pentagram represents the 3 × 3 convolution kernel, and the dark gray area containing the pentagram is the convolved receptive field. Figure 4(a) represents a 3 × 3 ordinary convolution, which has a receptive field size of only 3 × 3. Figure 4(b) represents a dilated convolution with Rate = 2, which can cover a 5 × 5 receptive field without increasing the number of parameters and computation and keeping the convolution kernel of 3 × 3 size.

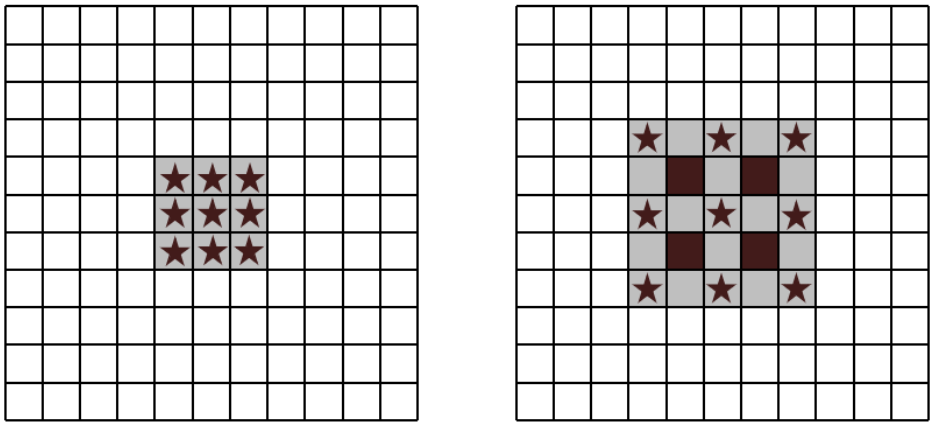

Figure 4: Ordinary convolution vs. dilated convolution

3.3 Decoder

Decoder is the key module in MCSnet for improving segmentation boundaries and details. Its main task is to fuse some of the shallow features of the encoder and gradually recover the feature map resolution by bilinear interpolation and refine the segmentation boundary [5].

Firstly, the feature map output from the encoder needs to be up-sampled 4 times to increase the resolution of the feature map output from the encoder, so that it can be close to the resolution of the original image, so that the details in the image can be effectively restored.

Next, the feature maps that have been upsampled need to be fused with the low-level feature maps extracted from the backbone network. Although the feature maps output from the encoder are rich in semantic information, they are relatively lacking in spatial information (points, lines, and surfaces), which may result in the segmented lesion boundaries not being accurate enough. The low-level feature maps, on the other hand, usually contain rich spatial information, which can help the decoder better recover the details in the image. However, due to the high number of channels in the low-level feature layer, if their importance is too high, it will affect the learning and training effect of high-level semantic features. Therefore, 1×1 convolution is used to compress the number of channels so that it is consistent with the proportion of channels occupied by the feature map obtained by the encoder after up-sampling, which retains the high-level semantic feature information well and compensates for the loss of details caused by the down-sampling process.

Finally, by subjecting the fused feature maps to 3×3 convolution and 4-fold upsampling, the number of channels in the feature maps can be effectively adjusted to be consistent with the original image, recovering the original spatial information of the image. In addition, this can make the boundary of the target finer, thus obtaining the segmentation prediction results.

ISSN 2616-5775 Vol. 7, Issue 9: 60-66, DOI: 10.25236/AJCIS.2024.070908

4. Experimental analysis

4.1 Evaluation metrics

When evaluating a typical deep learning model, the predictions are usually classified into four categories: True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN) [6]. Where TP denotes the number of correctly identified labeled pixels in the segmentation result, FP denotes the number of identified non-target pixels, and FN denotes the number of target pixels that are not identified.

In this paper, we mainly select two evaluation indexes, Mean Intersection over Union ($mIoU$) and Accuracy, to measure the segmentation effect of the network.

In image segmentation, Accuracy denotes the accuracy of an image segmentation algorithm in dividing an image into different parts. Specifically, Accuracy is the ratio of the size of the part of the image obtained after segmentation that is the same as the real image to the total size of the real image. Where Pixel Accuracy is the ratio of the number of pixels in the part of the image obtained after segmentation that is the same as the real image to the total number of pixels in the image obtained after segmentation; and $mIoU$ is the average of the number of pixels in the part of the image obtained after segmentation that is the same as the real image to the number of pixels in the same part of the image obtained after segmentation and the real image. The expressions are:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \tag{2}$$

$$mIoU = \frac{1}{k+1}\sum_{i=0}^{k} \frac{TP}{FN+FP+TP} \tag{3}$$

4.2 Analysis of $\mathbf{mIoU}$ results for different networks

As can be seen from Table. 2, the model MCSnet in this paper improves 3.13% compared to FCN, FCN fuses features by summing up the corresponding pixel values of the feature map, and although the use of jump connections improves the roughness of upsampling, there is still a certain inaccuracy in segmentation of the boundaries; compared to Unet, it improves 1.42%, although Unet uses dimensional splicing for feature fusion, there is still room for improvement in detail processing; compared with Deeplabv2, Deeplabv2 improves 3.39%, Deeplabv2 adopt4s pyramid pooling to deal with features of different scales, but there are some limitations in dealing with complex scenes, in pyramid pooling, the features of different scales are processed and aggregated separately, and in the process of feature fusion, some subtle differences in the features are blurred or lost.

Table 2: Comparison of different model $mIoU$

Model	$mIoU$(%)
FCN	83.20
Unet	84.91
Deeplabv2	82.94
MCSnet	86.33

In Figure 5, the $mIoU$ comparison between four different semantic segmentation models (MCSnet, FCN, Unet, and Deeplabv2) is shown, where the horizontal coordinates indicate the number of trainings. It is clear from the figure that the MCSnet proposed in this paper performs best in terms of $mIoU$, followed by Unet and FCN, while Deeplabv2 is at the lowest level.

The smooth curves of the MCSnet proposed in this paper reflect its stability and robustness under different parameter settings. In contrast, Deeplabv2's folds are not smooth enough indicating its higher sensitivity to the dataset, resulting in higher performance fluctuations. While Unet is slightly lower than MCSnet in $mIoU$, its smooth curve indicates the model's relatively good stability in different situations.FCN reflects the possible faster convergence speed at the beginning of the training, and the final $mIoU$ level achieved is relatively high, which is related to the fact that the model efficiently captures more image features during the learning process.

Taken together, the result that MCSnet performs best in terms of $mIoU$ may stem from the fact that it employs a more advanced deep learning technique and a more complex structure that is better able to capture semantic information in the image. Unet and FCN, although they perform slightly less well than Deeplabv2, they have more advantages in terms of stability and convergence speed. Deeplabv2, on the

ISSN 2616-5775 Vol. 7, Issue 9: 60-66, DOI: 10.25236/AJCIS.2024.070908

other hand, needs further tuning and optimization to achieve a better level of performance.

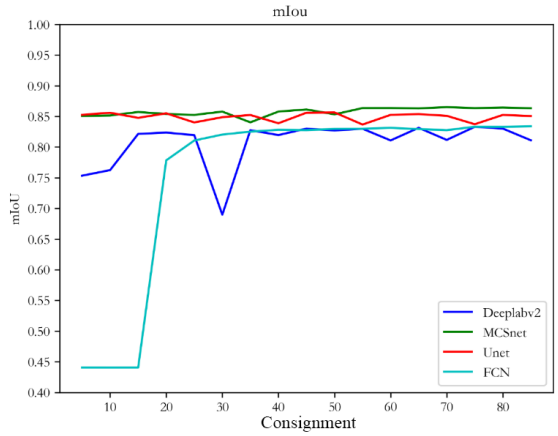

Figure 5: mIoU comparison chart

4.3 Analysis of segmentation results

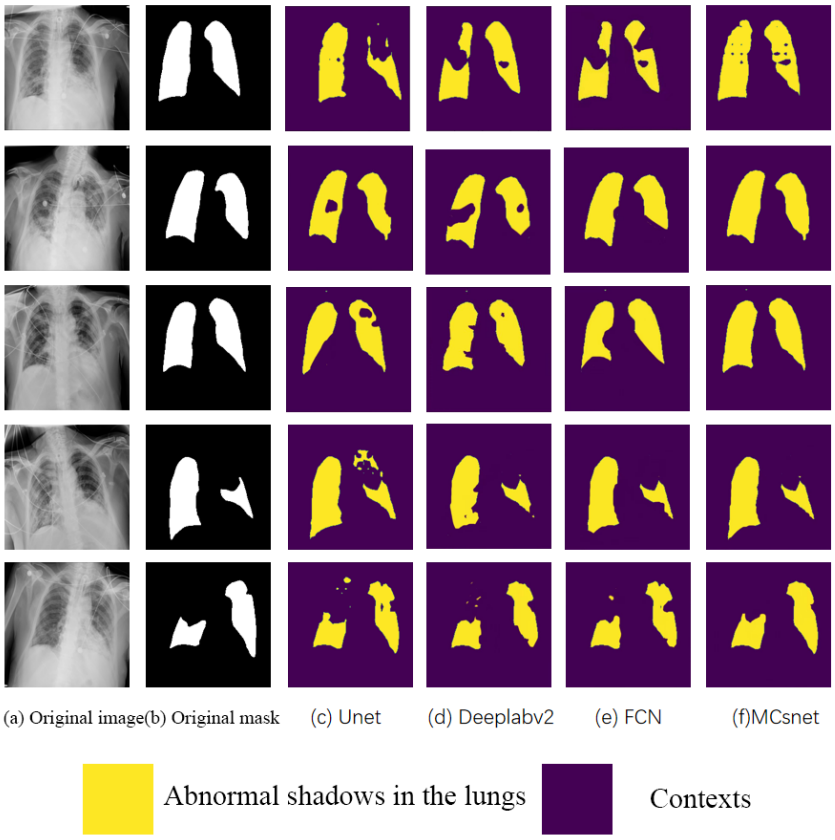

Figure 6: Comparison of segmentation effect of different models

Some of the segmentation results of Unet, Deeplabv2, FCN and MCSnet semantic segmentation models are shown in Figure 6. In the segmentation results of Unet network, the segmentation accuracy is low, and the segmentation boundary is not clear, which is related to the network structure of Unet. The outputs of each layer of the encoder of the Unet network are directly connected to the decoder at the same layer, and the model can not take full advantage of the deep and shallow semantic information, which makes the model semantic segmentation accuracy lower, which makes the model semantic segmentation accuracy lower. The pyramid pooling module of Deeplabv2 can capture multi-scale information in the image, but it is easy to lose the detailed information of the image, which indicates that the encoder feature

ISSN 2616-5775 Vol. 7, Issue 9: 60-66, DOI: 10.25236/AJCIS.2024.070908

extraction ability of this network is poor. The segmentation effect of the FCN network is acceptable because it can capture more contextual information and up-sampling through the inverse convolutional layer, which maintains a higher resolution. The MCSnet network segmentation is the most effective, and by utilizing multi-scale contextual information, the semantic information on different scales in the image can be better understood, which improves the segmentation accuracy for fine structures and boundaries. Therefore, it shows good performance in segmentation of X-ray pneumonia images.

5. Conclusions

In this study, we demonstrated significant advantages in pneumonia diagnosis through the proposed Multi-scale Cascade Segmentation Network (MCSnet), which combines an encoder, an ASPP module, and a decoder to efficiently extract multi-scale semantic information for accurate segmentation and abnormality recognition of lung radiographs. Compared with traditional methods, MCSnet exhibits higher accuracy and $mIoU$ on the qata_v2 pneumonia detection dataset, providing reliable support for pneumonia diagnosis. With this study, we demonstrated the potential of deep learning techniques in medical image analysis, providing new possibilities for improving the efficiency and accuracy of pneumonia diagnosis. This study provides a useful reference for the further development of intelligent medical image diagnosis technology in the future and opens up a new path for the application of artificial intelligence in the medical field.

References

[1] Mei X, Lee H C, Diao K, et al. Artificial intelligence–enabled rapid diagnosis of patients with COVID-19[J]. Nature medicine, 2020, 26(8): 1224-1228.
[2] Shorten C, Khoshgoftaar T M. A survey on image data augmentation for deep learning [J]. Journal of big data, 2019, 6(1): 1-48.
[3] Oliveira G L. Encoder-decoder methods for semantic segmentation: Efficiency and robustness aspects [D]. Albert-Ludwigs-Universität Freiburg, 2019.
[4] Li Y, Li K, Chen C, et al. Modeling temporal patterns with dilated convolutions for time-series forecasting [J]. ACM Transactions on Knowledge Discovery from Data (TKDD), 2021, 16(1): 1-22.
[5] Wang K, Yang J, Yuan S, et al. A lightweight network with attention decoder for real-time semantic segmentation [J]. The Visual Computer, 2022, 38(7): 2329-2339.
[6] Tharwat A. Classification assessment methods [J]. Applied computing and informatics, 2021, 17(1): 168-192.