Spine Image Segmentation Based on U-Net and Atrous spatial pyramid pooling

Zhen Yang1 a*, Long Chen1 b*, Ting Fu1 c*, Zhijian Yin1 d*, Fan Yang1 e*

1 Jiangxi Science and Technology Normal University, Nanchang 330013, China
a*yangzhenphd@aliyun.com, b*1020170701@jxstnu.com, c*criafting@aliyun.com, d*zhijianyin@aliyun.com, e*kooyang@aliyun.com

Abstract—In the past ten years, deep learning has achieved remarkable results in the area of natural image segmentation, and has gradually turned to the field of medical image segmentation. The precise segmentation of spine images can be used for early screening of spondylopathy, which is convenient for early detection and treatment of patients. Aiming at the segmentation of the spine image by the U-Net network, which structure will lead to large model calculation, network overfitting, image size, noise information and other issues. This paper introduces a new method based on spatial pyramid pooling module ASpp on U-Net network and applies the proposed network structure to the segmentation of the spine image. The model is first enhanced by the Spp module and Densenet data; secondly, the U-Net encoding and decoding structure is adopted, and the deep separable convolution is used. This method greatly reduces the complexity and computation of the model, and uses a rectangular convolution kernel to increase the computation of the model in a small amount. Based on the volume, the receptive field of the convolution operation is expanded; finally, in order to effectively enhance the segmentation area of the image, the SE-Net attention mechanism module is added to the lateral connection. The method proposed in this paper conducts ablation experiments on a public dataset of spine images and compares it with the U-Net network. The method proposed IOU in this paper can reach 72%, and the F1-score can reach 0.92. The comparison of the experimental results of spine medical image datasets shows that the method proposed in this paper can effectively improve the accuracy and accuracy of spine image segmentation.

1. Introduction

In the last decade, deep convolutional neural networks (CNNs) have been widely used in computer vision community, and have achieved great progress in a board range of tasks. e.g., image classification, object detection and instance segmentation[1]. Results of Deep Learning in Computer Vision was inspired by a lot of scholars, and applied it to medical image segmentation. However, the instance segmentation is still a challenging task in the computer vision, which requires an algorithm to predict a per-pixel mask with a category label for each instance of interest in an image[2]. U-Net network structure as the earliest deep learning medical image segmentation framework. However, the U-Net network loses the feature information of medical images when the encoder and decoder are connected laterally, resulting in a feature semantic gap [3].

The U-Net network is a fully convolutional neural network[4]. The input and output are images without fully connected layer. The shallower high-resolution layer is used to solve the problem of pixel positioning, and the deeper layer is used to solve the problem of pixel classification[5]. However, the U-Net still has some disadvantages and shortcomings, such as: effective convolution increases the
difficulty, universality of model design, and the form of trimming is not symmetrical with the Feature Map[6].

Therefore, to overcome these problems in the process of image feature extraction, We propose a scheme for implementing image instance segmentation. The main strategies of this paper include the following three points[7]:

1) Inputting image through the Spp module. After the image passes through the Spp module, the final output image is the scale[8];

2) In the U-Net network, Atrous convolution is used to replace the original ordinary 3x3 convolution, which can expand the image receptive field without losing resolution. In this way, the U-Net network can simultaneously pay attention to the context information of the spine image area, and obtain more high-level semantic features of the image;

3) The attention mechanism SE-Net is added after each horizontally connected block of the U-Net network, which increases the accuracy and effectiveness of the network model for the segmentation of the target area while increasing the lightweight computation.

In section 2, the paper will introduce recent work, and in section 3 the paper will introduce their framework. In the last, we will show our experimental process and results, at the same time we will summarize our work and future prospects.

2. Related Works
In the computer vision, deep learning plays a key role in the application of face recognition, object detection, object tracking, defect detection and so on[9]. The precise segmentation of spine medical images is significant for the diagnosis of diseases such as spondylitis and cervical spondylitis. Therefore, the precise segmentation of spine images can be used for early screening of some spinal diseases, facilitating the early detection and treatment of patients.

Dilated convolution can expand the receptive field without loss of feature image resolution, which is very useful in target detection and segmentation tasks[10]. On the one hand, dilated convolution can detect and segment large targets in a large field, on the other hand, the high resolution can accurately locate target. Moreover, when different dilation rates are set for the dilated convolution, the receptive field will be different, that is, multi-scale information is obtained[11]. In the field of computer vision, multi-scale image information plays an important role in vision tasks.

One of the representative methods is squeeze-and-excitation network, which learns channel attention for each convolution block and provides a significant performance improvement for various deep CNN architectures[12].

Due to the variability of spine images, there are differences in size, brightness, and shape between patients, which makes it impossible to judge the boundaries. Therefore, the segmentation of many spinal images in hospitals mainly relies on the manual annotation of doctors and medical experts to segment the foreground and background information of spinal images. Yet, With the development of society, the number of patients is also increasing, manual labeling not only takes time and effort, but also has a high rate of misjudgment, which affects doctors’ follow-up treatment methods. The method we propose in this paper improves the U-Net network, which can effectively segment medical spine images and greatly improve the accuracy and accuracy of segmentation.

3. Methods
The method proposed in this paper is mainly composed of four parts: (1) Spatial Pyramid Pooling; (2) U-Net; (3) ASp; (4) SE-Net module. In this section, we will describe our overall framework structure in detail, as shown in Figure 1, and describe the details of each branch in detail.
Figure 1. The network model structure of our method

1) Spatial Pyramid Pooling

As shown in Figure 2, the black picture represents the feature map after convolution, then we extract the features of the feature map with three blocks of different sizes, which are $4 \times 4$, $2 \times 2$, and $1 \times 1$, and put these three grids below. Through feature extraction of three blocks of different sizes, on the input image feature map, we can get $16+4+1=21$ different blocks (Spatial bins). From these 21 different blocks, we extract a feature from each block, which is exactly the 21 dimensions we want to extract. This process of pooling by combining different sizes of grids is called spatial pyramid pooling (Spp). For example, the maximum pooling of spatial pyramid is to calculate the maximum value of each block from 21 image blocks to obtain an output unit, and finally obtain a 21-dimensional feature output. In this way, the required number of features can always be obtained for different inputs, without fixing the input size of the convolution layer.

Figure 2. The network model structure of our spatial pyramid pooling

2) U-Net

As shown in Figure 3, in the structure of U-net, it includes a contraction path that captures context information and a symmetrical expansion path that allows precise positioning[13]. The U-Net network adopts an end-to-end training method, that is, the input is an image, and the output is also an image, and the best results can be obtained with very little data. The structure of U-Net is shown in the Figure 3. The left side of the U-Net network is the contraction path for encoding, and the right side is the expansion path for decoding. The encoder has four sub-modules, each containing two convolutional layers. After each sub-module, there is a down-sampling layer implemented by max pool[14]. The resolution of the input image is $572 \times 572$, and the resolutions of modules 1-5 are $572 \times 572$, $284 \times 284$, $140 \times 140$, $68 \times 68$ and $32 \times 32$ respectively. Since the convolution uses the valid mode, the resolution of the next
The sub-module here is equal to \((\text{resolution of the previous sub-module} - 4)/2\). The decoder consists of four blocks, and through the up-sampling operation, the feature image resolution is restored in turn (since the convolution uses valid mode, the actual output is smaller than the input image). The U-Net network also uses lateral connections to fuse the features of the encoder and decoder, and fuse the results of the encoder network up-sampling and the decoder up-sampled results as the input of the next block.

![Diagram](image)

**Figure 3.** The network model structure of U-Net

3) ASpp

ASpp was proposed in deeplabv2. In the semantic segmentation task, we hope to have a larger receptive field for the features extracted from the images, and hope that the resolution of feature images will not decrease too much (too much resolution loss will lose a lot of detailed information about the image boundary), but these two are contradictory. If larger receptive fields are to be obtained, larger convolution kernels or larger steps are needed for pooling. For the former, the amount of calculation is too large, and the latter will lose resolution. Moreover, the dilated convolution is used to solve this contradiction, and it can get a larger receptive field without losing too much resolution. The dilated convolution is shown in Figure 4. Compared with ordinary convolution, atrous convolution enlarges the receptive field when the pooling loss information is small, so that each convolution output contains a wider range of information, ensuring that the output image features have a higher resolution. As shown in Figure 4, the ASpp module uses parallel convolution layers of holes with different dilation rates to capture multi-scale information on feature images, and fuse the output results to obtain image segmentation results. As shown in the Figure below, this is an improved ASpp in deeplab v3. A 1×1 convolution and three 3×3 dilated convolutions are used, and each convolution kernel has 256 channels and a BN layer. In fact, the 1×1 convolution is equivalent to a dilated convolution with a large rate, because the larger the rate, the smaller the effective parameters of the convolution kernel. Therefore, this 1×1 convolution kernel is equivalent to the large rate convolution kernel.
4) SE-Net Attention

As shown in Figure 5, in the network structure of SE-Net, given an input x, the number of feature channels is C1. After convolution, compression, decompression, connection and other operations, the number of feature channels is finally obtained as C2. The SE-Net network first squeezes the operation, which uses a global pooling operation to synthesize the input features of size H*W*C into a feature description of C*1*1. Turn each two-dimensional feature channel into a real number with a global receptive field, and match the output dimension H*W*C with the input feature channel number. Second, the Excitation operation, which is a gate mechanism in a recurrent neural network, generates weights for each channel in the feature map by optimizing parameters, and sets different weights for different parts of the image, where the parameters are learned to explicitly construct features Correlation between channels. Finally, the Reweight operation weights the weight of the excitation output to the previous feature channel by channel by multiplication, Complete the re-calibration of the original features in the channel dimension and use them as the input data for the next level.

4. Experiment

4.1. Implementation Details

The deep learning framework used in this experiment is Keras, which is based on the Tensorflow2.0 backend. The hardware configuration for network training is Intel Core i7-10700CPU@2.9GHz, GeForce GTX 1660Ti, Nvidia driver-450, Ubuntu18.04.5 LTS 64bit operating system, 16G RAM, 2.2TB Disk, and python3.6 and CUDA10 software configuration.0CUDNN7.0.5, image saving and other processing are based on the open-source software OpenCV.

In the model proposed in this paper, the size of the input image is uniformly 512*512*3, and in order to prevent distortion, the image is fine-tuned. The input image preprocessing input is normalized by dividing it by 255. When training the network, the batch_size is set to 4, and customized data enhancement methods were used, such as random direction translation, random contrast enhancement, random gamma transformation and random small-scale scaling, etc; each iteration steps_per_epoch is 50; the training epochs is 300 rounds; the early stopping method is used to control the training process. If the loss value of the verification does not decrease within 10 epochs, the training will be stopped to
prevent overfitting. The optimizer is Adam, and the initial learning rate is both 0.0001; the learning rate training strategy is that if the loss does not decrease within 3 epochs during the training process, the learning rate is reduced to 0.1 times the original.

In this paper, the experimental data uses the public medical spine dataset, which is a spine image randomly selected from different ages and human bodies. The training set contains 1211 spinal images and corresponding masks; the remainder is in the test set. Real pixel labels are used for testing. The spine image is in PNG format, the size of each image is 20kb, the resolution of each spine image is 120 pixels wide and 120 pixels high. The images and labels are shown in Figure 6:

![Spine image and annotation information](image)

The input image is uniformly fine-tuned to 512*512*3 size. The input image preprocessing input is divided by 255 for normalization. The normalization formula of $x_{\text{norm}} = \frac{x - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}}$ is as follows:

$$x_{\text{norm}} = \frac{x - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}}$$

### 4.2. Evaluation Metrics

- $\text{IoU} = \frac{T_P}{T_P + F_N}$
- $\text{Rcall} = \frac{T_P}{T_P + F_N}$
- $\text{Rprecision} = \frac{T_P}{T_P + F_P}$
- $F_1 = \frac{2 \cdot \text{Rprecision} \cdot \text{Rrecall}}{\text{Rprecision} + \text{Rrecall}}$

In the above formula, $T_P$ represents the number of positive samples tested, and the test is correct, that is, true positive; $T_N$ represents the number of negative samples tested, and the test is correct, that is true negative; $F_P$ represents the number of positive samples tested, but the detection is wrong, that is false positive; $F_N$ represents the number of negative samples detected, but the detection is wrong, that is false negative.

### 4.3. Loss Function

Our method uses a loss function with boundary weights:

$$\text{Loss}_{\text{Dice}} = 1 - \frac{2 |X \cap Y|}{|X| + |Y|}$$

$X$ is the Ground Truth segmented image, and $Y$ is the predicted segmented image. Dice coefficient is a key evaluation index in medical image segmentation effect. The Dice loss function compares the intersection and union between the prediction result area and the ground truth area, and calculates the function loss by taking all the pixels in a category as a whole.

$$L = -y \log y - (1 - y) \log (1 - y') = \begin{cases} y & y = 1 \\ - \log (1 - y') & y = 0 \end{cases}$$

Among them, $\gamma$ is the key factor to increase the loss of cross entropy. When $\gamma > 0$, the loss of classified samples can be effectively reduced, and the problem that difficult samples are difficult to
mine can be solved. $\alpha$ is a balance factor used to balance the problem of unbalanced proportion of positive and negative samples.

4.4. Experimental images and forecast results

![Model Loss](image1)

![Model Metrics](image2)

| Algorithm   | IOU/% | Recall/% | F1-score | Dice/% | Sensitivity/% | Specificity/% |
|-------------|-------|----------|----------|--------|---------------|---------------|
| U-Net       | 65.10 | 87.84    | 0.81     | 78.36  | 87.84         | 94.66         |
| SE-U-Net    | 66.82 | 88.32    | 0.84     | 79.74  | 88.49         | 95.56         |
| Our Method  | 72.28 | 90.87    | 0.90     | 84.77  | 89.13         | 97.18         |

**Figure 7.** Comparison results on three different methods

**Table 1.** A table with headings spanning two columns and containing notes

According to the comparison of ablation experimental results, the method proposed in this paper is the best for medical image segmentation, and our experimental results have obvious improvement in F1-score and IOU, which shows that ASpp and SE-Net network have some advantages in U-Net spine image segmentation.

4.5. Experimental images and forecast results
5. Conclusion
Aiming at the insufficient extraction of medical image features by the ordinary convolutional neural network in the U-Net network, the ASpp-UNet network proposed in this paper adds the Spp module to the U-Net network, and replaces the ordinary convolution with the hole convolution and the cavity convolution and the SE-Net attention mechanism to have stronger feature extraction for various complex spine images Ability and only a small number of calculations have been added. The public data set of spine images shows that the experiment has effectively increased the IOU and F1-score, and it can also effectively extract the key information of the spine image in the image. For the segmentation of spine images, this paper replaces the convolutional network with the ASpp network on the basis of the U-Net network model, and adds the attention mechanism SE-Net. Ablation experiment results show that our proposed new network structure can efficiently segment spine images and obtain higher segmentation results. Deep learning image segmentation is also a research direction that people will pay attention to.

Acknowledgement
This work was supported by the National Natural Science Foundation of China (61866016, 62061019), General Project of Jiangxi Natural Science Foundation (20202BABL202014), the Key Project of Jiangxi Education Department (GJJ201107), the Science Foundation of Artificial Intelligence and Bioinformatics Cognitive Research Base Fund of Jiangxi Science and Technology Normal University of China (2017ZDPYJD005).

References
[1] Wang Q, Wu B, Zhu P, et al. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks[J]. 2019.
[2] Tian Z, Shen C, Chen H. Conditional Convolutions for Instance Segmentation[M]. 2020.
[3] Zhou Z, Siddiquee M, Tajbakhsh N, et al. U-Net++: A Nested U-Net Architecture for Medical Image Segmentation[C]// 4th Deep Learning in Medical Image Analysis (DLMIA) Workshop. 2018.
[4] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation[J]. Springer International Publishing, 2015.
[5] X. Duan, K. Jia, B. Li, D. Guo, E. Zhang and C. Qin, "Reversible Image Steganography Scheme Based on a U-Net Structure," in IEEE Access, vol. 7, pp. 9314-9323, 2019, doi: 10.1109/ACCESS.2019.2891247.
[6] Zhang M, Yu F, Zhao J, et al. BEFD: Boundary Enhancement and Feature Denoising for Vessel Segmentation[J]. 2021.
[7] H. Wen, L. Chen, T. Fu, Z. Yang and Z. Yin, "Detecting the Surface Defects of the Magnetic-Tile Based on Improved YOLOACT++," 2021 IEEE 21st International Conference on Communication Technology (ICCT), 2021, pp. 1097-1102, doi: 10.1109/ICCT52962.2021.9658066.
[8] He K, Zhang X, Ren S, et al. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2014, 37(9):1904-16.

[9] Lei Junming, Wang Chenggang, Liu Xueyan. Retinal Vessel Image Segmentation Based on Improved U-Net[J]. Information Technology and Information Technology, 2019(10):168-169.

[10] Long J, Shelhamer E, Darrell T. Fully Convolutional Networks for Semantic Segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 39(4):640-651.

[11] F Yu, Koltun V. Multi-Scale Context Aggregation by Dilated Convolutions[J]. 2016.

[12] Jie H, Li S, Gang S, et al. Squeeze-and-Excitation Networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, PP(99).

[13] Weng W, Zhu X. INet: Convolutional Networks for Biomedical Image Segmentation[J]. IEEE Access, 2021, PP(99):1-1.

[14] H. Wu, Y. Zhang, W. Fu, C. Zhang and S. Niu, "A Novel Pre-Processing Method for Neural Network-Based Magnetic Field Approximation," in IEEE Transactions on Magnetics, vol. 57, no. 10, pp. 1-9, Oct. 2021, Art no. 9200109, doi: 10.1109/TMAG.2021.3095590.