Hindawi
Security and Communication Networks
Volume 2021, Article ID 2139024, 8 pages
https://doi.org/10.1155/2021/2139024

Research Article

J-Net: Asymmetric Encoder-Decoder for Medical Semantic Segmentation

Yanli Shi [1] and Pengpeng Sheng [2]

[1] *School of Science, Jilin Institute of Chemical Technology, Jilin City 132022, China*
[2] *College of Information and Control Engineering, Jilin Institute of Chemical Technology, Jilin City 132022, China*

Correspondence should be addressed to Yanli Shi; syl@jlict.edu.cn

Received 17 May 2021; Revised 17 July 2021; Accepted 18 August 2021; Published 31 August 2021

Academic Editor: Feiran Huang

With the development of deep learning, breakthroughs have been made in the field of semantic segmentation. However, it is difficult to generate a fine mask on the same medical images because medical images have low contrast, high resolution, and insufficient semantic information. In most scenarios, existing approaches mostly use a pooling layer to reduce the resolution of feature maps. Therefore, it is difficult for them to consider the whole image features, resulting in information loss and performance degradation. In this paper, a multiscale asymmetric encoder-decoder semantic segmentation network is proposed. The network consists of two parts, which perform feature extraction and image restoration on the input, respectively. The encoder network obtains multiscale feature information by connecting multiple ASPP modules to form a feature pyramid. Meanwhile, the upsampling layer of each decoder can be connected to the feature map generated by the corresponding ASPP module. Finally, the classification information of each pixel is obtained through the sigmoid function. The performance of the proposed method can be verified on publicly available datasets. The experimental evidence shows that the proposed method can take full advantage of multiscale feature information and achieve superior performance with less inference computational cost.

1. Introduction

Since the proposal of AlexNet [1] by Professor Hinton in 2012, computer vision has made breakthroughs in image classification, target detection [2, 3], and semantic segmentation [4]. Alternatively, computer-aided diagnosis systems based on neural network convolution have been widely used in many medical image analysis tasks.

Segmentation is one of the most important and popular tasks in medical image analysis. It plays a vital role in disease diagnosis, surgical planning, and prognostic evaluation. We urgently should improve the efficiency of doctors' diagnoses to save patients' lives. Medical image segmentation methods and theories are many, including borders, thresholds, regional growth, statistics, graph theory, active contour, information theory, fuzzy set theory, and neural network. Due to the large hints of computing power, neural networks have gradually become the preferred technology for semantic segmentation tasks and have achieved excellent results in various competitions.

From [5, 6], it can be seen that the features of the bottom layer (such as the output of layers 1, 2, and 3) are more biased toward the basic units of the image, such as points, lines, and edge contours, while the high-level semantic features are layers 4 and 5. It is more abstract and more similar to the semantic information of the image, more like a region. Based on the above understanding, the focus of the semantic segmentation network is how to better combine high-level semantic information with low-level feature information. In medical image segmentation tasks, FCN [6] and U-Net [7] are the mainstream network models. The other architectures [8–12] use different mechanisms (long jump connection, pyramid pooling, and so on) as part of the decoding mechanism. The difference between these architectures lies mainly in the decoder network. The decoder has the task of the encoder to learn distinguishable characteristics (lower resolution) of semantic mapping pixel space (higher resolution) to obtain a dense classification. Semantic segmentation has the ability to not only distinguish the pixel level but also learn the characteristics of the different stages required in the encoder.

In some recent work, RSANet [13] employs residual semantic-guided attention mechanism (RSAM) to fuse the multiscale features from LCNet for improving detection performance efficiently. In ALNet [14], the encoder adopts a novel residual module to abstract feature representations. Swin transformer [15] introduced transformer in semantic segmentation to increase the model's ability to capture long-distance information.

To reduce the computational complexity and improve the training speed, the traditional encoder-decoder must use many downsampling processes, which will cause some small features to be lost in the downsampling process, and it is impossible to accurately classify each pixel. To solve the problem of information loss, we propose a new multiscale fusion method of asymmetric encoder-decoder network (J-Net) to enhance the use of multiscale feature information, while shortening the information flow channel. The contributions of our works are threefold:

(i) An asymmetric encoder-decoder network is proposed to further reduce the loss of information in the downsampling process.

(ii) Proposing a new connection method between the encoder network and the decoder network can reduce unnecessary information flow.

(iii) Experimental results show that our framework better integrates features of different scales and achieves excellent performance with less computational cost.

The rest of this article is organized as follows. In Section 2, several backbone architectures used in modern semantic segmentation are reviewed. In Section 3, the J-Net structure and its concept are proposed. Section 4 compares FCN [4], DeepLab [16–19], and U-Net [7] and verifies the effectiveness of J-Net. Section 5 summarizes the advantages and disadvantages of J-Net in other fields.

2. Related Work

Since the proposal of FCN in 2015 [4], convolutional neural network has made considerable progress in the field of image segmentation. It has been shown that the main factor affecting image segmentation is how to expand the local receptive field and keep the loss of features in the process of downsampling. There are some mainstream technologies to obtain global information as follows (see Figure 1 for illustration).

2.1. Upsampling. The maximum change of FCN [4] compared with classification neural networks is that the classification network will add some convolutional layer at the end of the network so that a two-dimensional feature map can be obtained, followed by softmax to obtain the classification information of each pixel. This is the beginning of using convolutional neural networks to solve semantic segmentation problems. But in this way, direct upsampling only uses high-level semantic information and ignores low-level features, which affects the segmentation effect.

2.2. Encoder-Decoder. The encoder-decoder is a concept in the NLP field, not a specific algorithm, but a framework to solve problems. The model consists of two parts: the encoder network (feature extractor) and the decoder network (generator). The image features are extracted by stacking multiple feature extraction blocks (Conv + BN + RELU), and the local receptive field is expanded by repeated downsampling to obtain larger global features. The function of the decoder network (generator) is to generate the high-level feature vector into the target vector. This method also has defects. For input with a large amount of information, the encoder process will lead to the loss of information.

To solve this problem, many researchers have made various attempts. For example, the U-Net [7] uses skip-connect operation, and the feature map of each convolution layer of U-Net will be concatenated to the corresponding upsampling layer. In SegNet [20], the decoder uses the pooled indexes calculated in the max-pooling process to calculate the nonlinear upsampling of the corresponding encoder. In addition to the above two methods, there are other variants, such as using fixed (sparse) index array to sample or use replication upsampling. However, these methods consume more memory, require a longer convergence time, and do not perform well.

2.3. Atrous Convolution. Multiple downsampling will lead to information loss, which will make the network miss smaller targets when performing detection tasks and affect the final results of the network. In DeepLab [16–19], atrous convolution has been proposed. There are two functions of atrous convolution: one is to control the receptive field and the other is to adjust the resolution. Firstly, by adjusting the hole convolution rate, the receptive field in the center of the convolution core increases. Secondly, by setting the stride size, the hole convolution can increase the receptive field and reduce the resolution.

3. Method

To solve the problem of information loss caused by multiple downsampling, we propose an asymmetric multiscale encoder-decoder model. The proposed network is a modified FPN [21]. We change the way to get global features and reduce the proportion of encoders.

3.1. Network Architecture. The network architecture is shown in Figure 2. The multiscale feature fusion asymmetric encoder-decoder network consists of two parts: a larger encoder and a smaller decoder. This asymmetric structure

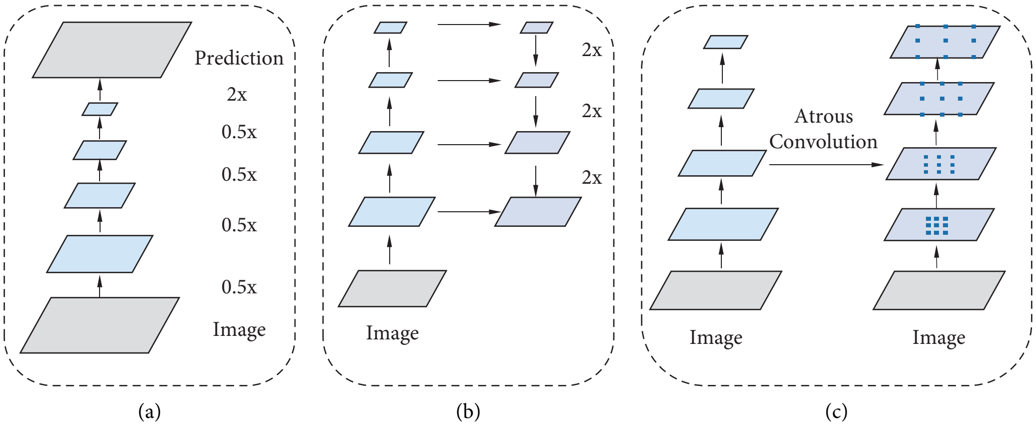


Figure 1: Alternative architectures to capture multiscale context. (a) Upsampling block. (b) Encoder-decoder block. (c) Atrous convolution block.

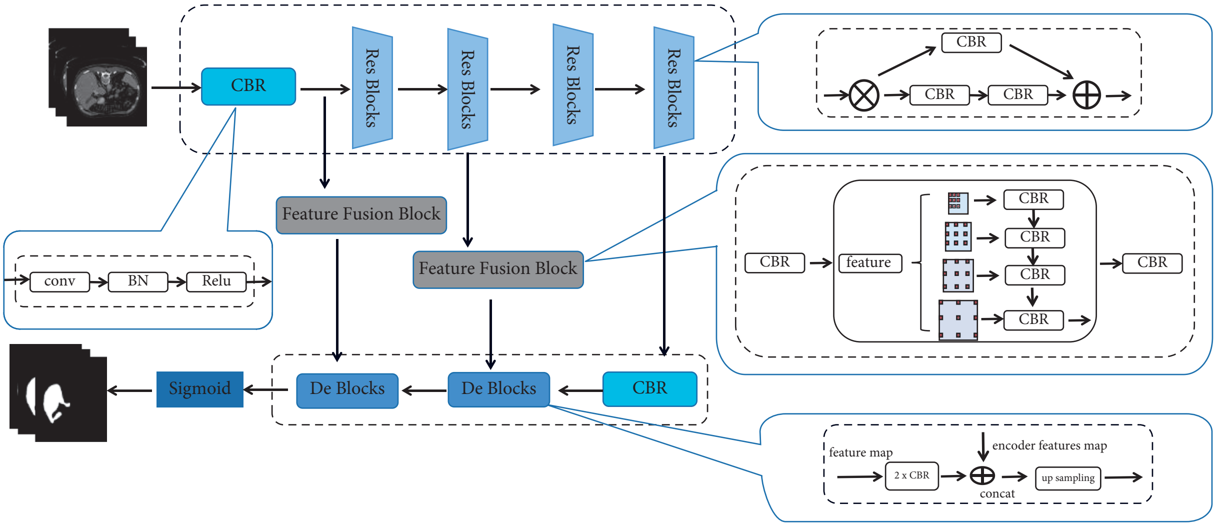


Figure 2: Network architecture.

reduces the redundancy of the network structure using ResNet as the encoder. At the same time, ASPP is performed on feature maps of different scales to obtain multiscale information, followed by softmax to obtain the classification information of each pixel. The feature extraction module consists of a 3 × 3 convolutional layer, BN layer, and RELU activation function layer. In the feature recovery module, the input is the feature pyramid generated by the ASPP of two feature maps, which are merged through the concatenated operation as an input to the next feature recovery module. The ASPP operation uses hole convolution to obtain feature pyramids on the same feature map and at the same time uses stride size to control the size of the output feature map unchanged, which is conducive to the fusion of multiscale features.

The encoder network uses hole convolution to obtain a feature map of a specific scale and sets different hole convolution rates to obtain a larger local receptive field without losing feature information. There is no need to maintain the same decoding stage size as the feature extraction stage because the encoder network no longer uses downsampling multiple times.

3.2. Asymmetric Encoder-Decoder. U-Net is one of the earliest algorithms using a full convolution network for semantic segmentation. The symmetrical encoder-decoder structure including the compressed path and extended path used in this paper was innovative at that time, and it affected the design of the following segmentation networks to a certain extent. The symmetrical structure is to fully integrate feature information of different scales. The network architecture is illustrated in Figure 3(b).

In conventional computer vision tasks, compressed image resolution is mainly achieved through a pooling layer or a convolutional layer (stride=2). Similarly, when we do semantic segmentation or target detection, the main purpose of the compressed path is to expand the local receptive field of the convolution kernel to obtain global information. We can use other methods to obtain global

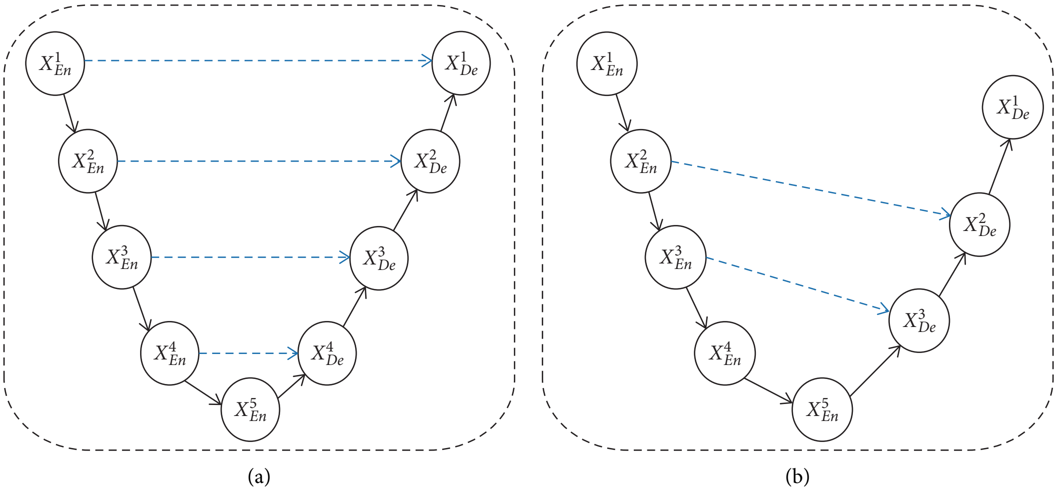


Figure 3: (a) Asymmetric encoder-decoder. (b) Symmetric encoder-decoder.

features, such as using a larger convolution kernel or hole convolution.

In this paper, we choose to stack multiple holes in convolution to reduce the use of downsampling. Using this method, the network can obtain a larger local receptive field without losing information, reduce the size of the decoder, and improve the training speed. The network architecture is illustrated in Figure 3(a).

The encoder is used as the feature extractor. Different from the use of VGG [22] as the backbone in FPN [21], here we use ResNet [23] as the backbone network. The use of a large number of skip-connect greatly improves the utilization rate of low-level features and can make the encoder network deeper. As a mask generator, the main purpose of the encoder is to restore the highly abstract feature vector to a mask image with the same size as the original image. In order to strengthen the high-level semantic features, we redesigned the encoder. The decoder consists of the repeated application of 3×3 convolutions (unpadded convolutions), each followed by a rectified linear unit (RELU) and a batch normalization unit (BN). Following the feature extractor, we use a dropout unit (Dropout) to prevent overfitting. In the final upsampling step, we restore the number of feature map sizes to the same as that of the input.

3.3. Multiscale Feature Map. In the field of object segmentation, one of the most basic principles is that the larger the receptive field of the final predicted pixel, the better the effect of capturing more contextual information and making more accurate predictions.

To obtain a larger receiving field, the mainstream method is to use large convolution kernels, hole convolution, and stack downsampling to reduce feature resolution so that convolution kernels of the same size can obtain larger local receptive fields.

For a large convolution kernel, due to its large size, it consumes several computing resources in network training. It is proposed in AlexNet [1] that multiple small convolution kernels can be connected in series to achieve the same receptive field as a large convolution kernel. However, the effective receptive field obtained using multiple stacked small convolution kernels is different from the theoretical receptive field [17]. Since these convolutions correspond to the difference between the effective receptive field and the theoretical receptive field, the feature information of the detected target is lost.

To solve this problem, we connect multiple atrous convolutions in parallel to form a spatial pyramid (ASPP) [16–19]. ASPP uses atrous convolutions with different dilation rates to perform different convolution operations on the feature map. It does not increase the number of parameters while obtaining a receptive field that exceeds the size of its convolution kernel. The most important thing is that the size of the feature map has not changed after the hole convolution operation. So, ASPP does not affect the original feature extraction operation of the encoder.

3.4. Feature Fusion. The feature pyramid is currently an important part of the target detection, semantic segmentation, behavior recognition, etc. It has excellent performance for improving model performance. References [9–12, 24–27] demonstrated various methods for constructing feature pyramids. The feature pyramid has a feature map of different scales. Targets of different sizes can have appropriate feature representations at the corresponding scales. By fusing multiscale information, targets of different sizes can be predicted at different scales. It improves the performance of the model very well. So, the most important thing to determine the mask is to obtain high-level semantic information (position, category, and so on) and low-level features (shape, color, and so on).

The serial use of ASPP allows us to obtain a richer view and combine feature information of different scales. We perform the atrous convolution operation on the advanced feature map to expand the range of the predicted receptive field of each pixel, combine intermediate features to constrain the approximate shape of the mask, and finally combine them with low-level features to standardize the edges of the mask. This multiscale feature fusion shortens the information flow path and at the same time increases the information flow path between the encoder and decoder and finally achieves the repeated use of important features.

4. Experiments

4.1. Datasets. The liver segmentation dataset has two sets: training set and test set, and the size of all images in the dataset is 512×512. The training set has 400 liver CT images and the corresponding segmentation template, and the verification set has a total of 20 liver CT images and the corresponding segmentation template. There are two categories of segmentation templates (liver and background).

The EM dataset has two sets: training set and test set, and the size of all images in the dataset is 512×512. The training set has 90 EM images and the corresponding segmentation template, and the verification set has 30 EM images and the corresponding segmentation template. There are two categories of segmentation templates (liver and background).

4.2. Implementation Details. We use the Adam algorithm as an optimizer, and the initial learning rate is set to 0.001. When using the gradient descent algorithm to optimize the objective function when getting closer and closer to the global minimum of the loss value, the learning rate should become smaller to make the model as close as possible to this point, and cosine annealing [28] can be achieved through the cosine function reduce the learning rate. The principle of cosine annealing is as follows:

$$\eta_t = \eta_{\min}^i + \frac{1}{2}\left(\eta_{\max}^i - \eta_{\min}^i\right)\left(1 + \cos\left(\frac{T_{\text{cur}}}{T_i}\pi\right)\right), \tag{1}$$

where i is the number of runs (index value); $\eta_{\max}$ and $\eta_{\min}$, respectively, represent the maximum and minimum values of the learning rate and define the range of the learning rate; T_{cur} indicates how many epochs are currently executed; and T_i indicates the total number of epochs in the $i-$ th run.

4.3. Experimental Results. To verify the effectiveness of the method in this paper, the traditional FCN, U-Net, and DeepLab networks were compared, the same data and parameter settings were used for training, and the trained model was verified with the test datasets.

For the comparative experiment [29, 30], we chose the Dice coefficient, precision coefficient, and recall coefficient as the evaluation criteria to measure the quality of the model. These evaluation criteria are as follows:

$$d = \frac{2\left(R_{\text{seg}} \cap R_{\text{gt}}\right)}{\left(R_{\text{seg}}\right) + \left(R_{\text{gt}}\right)}, \tag{2}$$

where the R_{seg} represents the predicted segmentation result and R_{gt} represents the segmentation result of ground truth. When applied to a binary segmentation task, it evaluates the degree of overlap between the predicted value R_{seg} and the true value R_{gt}.

Table 1: Segmentation results produced by different methods on the liver datasets.

Method	Backbone	Precision	Dice	Recall
FCN-8s	VGG16	0.8787	0.8537	0.8420
DeepLab-v3+	ResNet101	0.8773	0.8642	0.8605
U-Net	—	0.8774	**0.9172**	0.9622
J-Net	ResNet101	**0.8836**	0.9118	**0.9637**

Table 2: Segmentation results produced by different methods on the EM datasets.

Method	Dice
DeepLab-v3+	0.9185
FCN	0.9364
J-Net	**0.9376**

$$P = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{3}$$

where TP (true positive) represents predicting the positive class as a positive class number and FP (false positive) represents predicting the negative class as a positive class number. Precision indicates how many of the samples whose predictions are positive are truly positive samples.

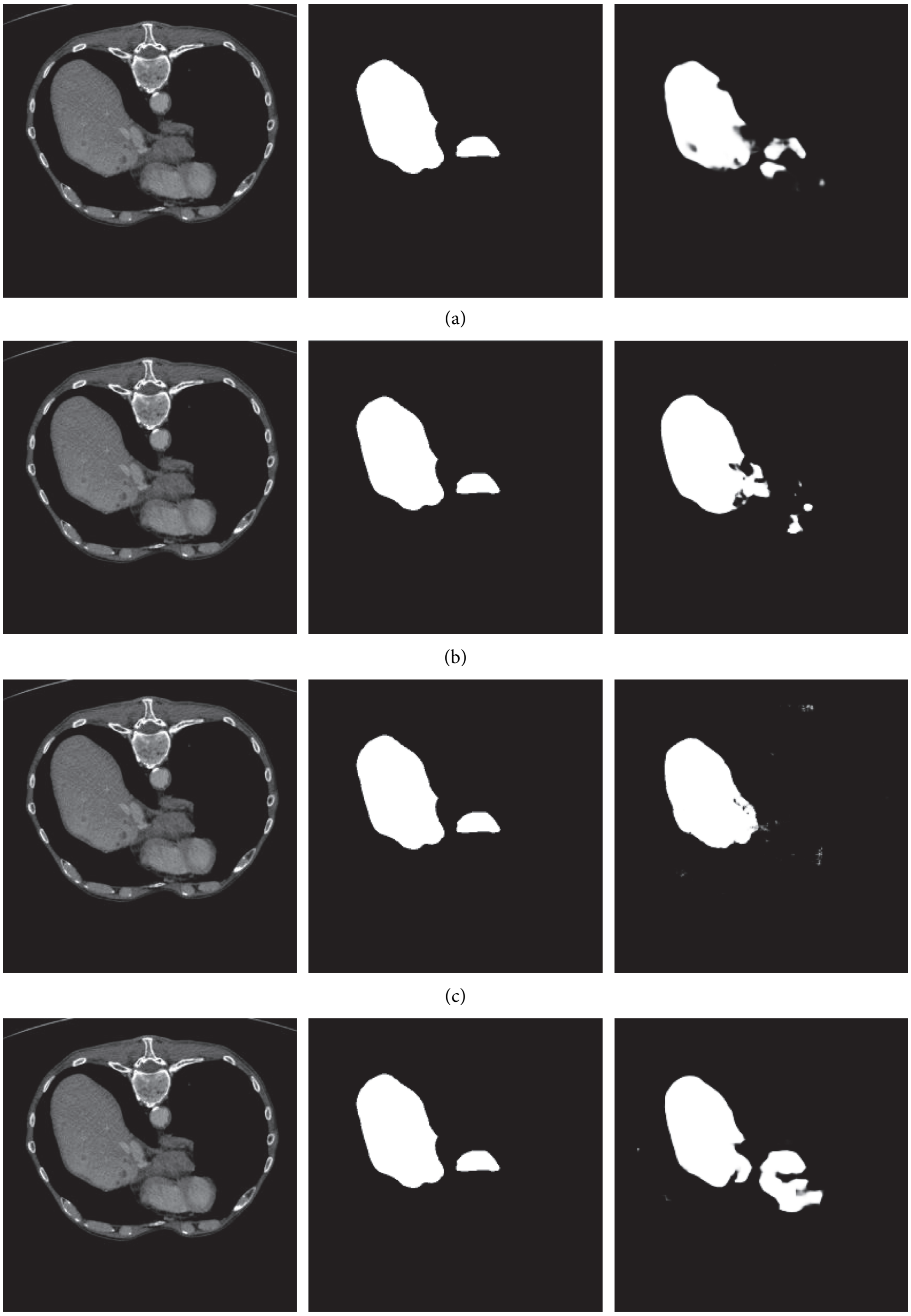
$$R = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{4}$$

where TP (true positive) represents predicting the positive class as a positive class number and FN (false negative) represents predicting the positive class as a negative class number. Recall rate indicates how many positive examples in the sample are predicted correctly.

For fair comparison, all the baselines are performed using the same hardware platform with a single NVIDIA GTX 3080 GPU. The minimum batch size is set to 4 (4 images per GPU); to stabilize the training at the beginning, the number of warm-up iterations has been extended from 30 to 50. Dice loss [12] is used as the loss function, Adam is used as the optimizer, the initial learning rate is 0.001, the minimum batch size is 4, and the epoch is 300. We need to consider overfitting when choosing an encoder network, so dropout [10] has been used to improve the generalization of the network. Tables 1 and 2 compare our J-Net with selected state-of-the-art networks.

The results in Table 1 indicate, for the liver dataset, that the encoder-decoder model is substantially more accurate than other segmentation models. We compare our method with FCN, DeepLab-v3+, and U-Net in liver segmentation tasks. The precision coefficient and recall coefficient of J-Net are 0.8836 and 0.9678 which are better than those of FCN, Deeplab-v3+, and U-Net, and J Net's DICE coefficient of 0.9129 is slightly lower than U-Net's DICE coefficient of 0.9172. However, the gap is not big. On the whole, the performance of J-Net is better than the that of above three algorithms. These all prove the effectiveness of our encoder-decoder architecture.

The segmentation results of liver dataset by different networks are shown in Figure 4. The figure shows the visual comparison on liver val set. From left to right are input


(a)

(b)

(c)

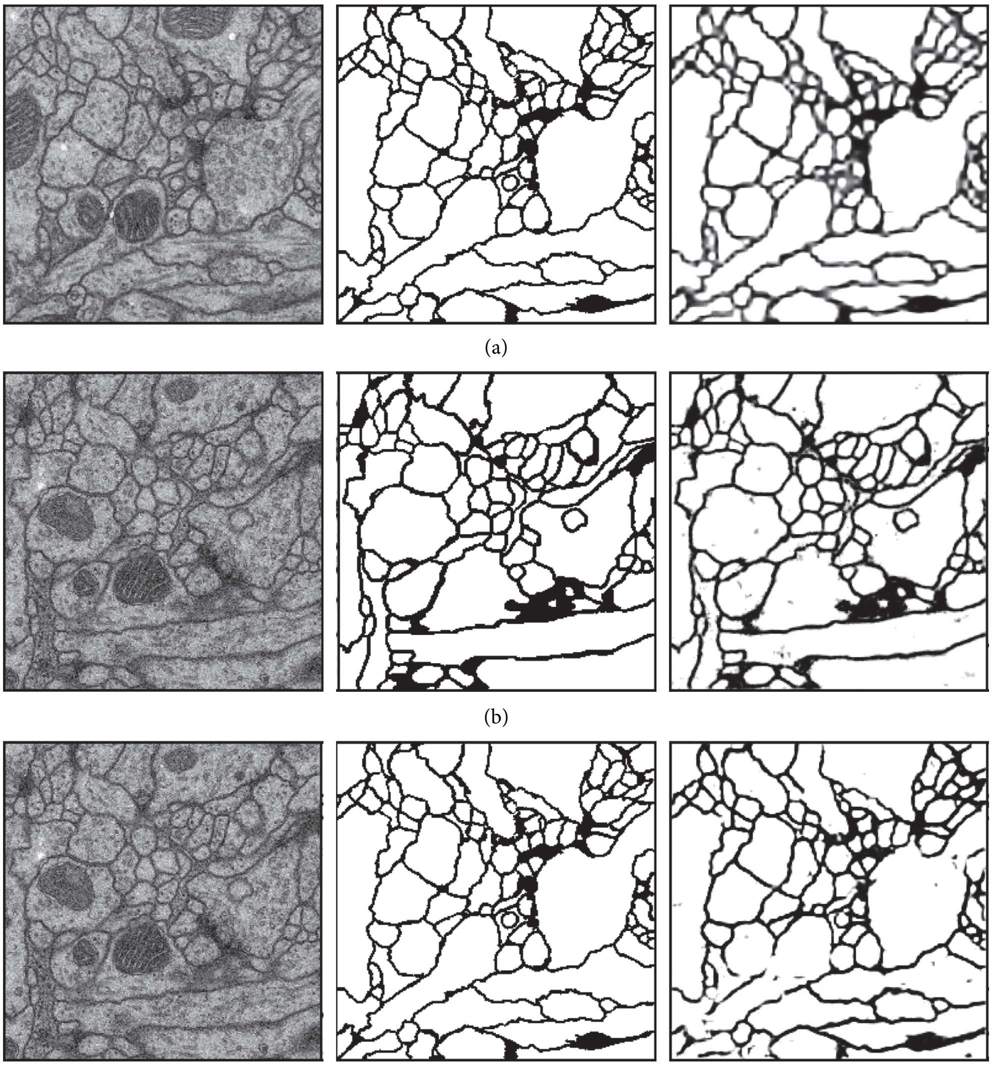
(d)

Figure 4: Segmentation results of liver datasets are compared among our J-Net and the other three proposed models. (a) J-Net. (b) U-Net. (c) FCN. (d) DeepLab-v3+.

images, ground truth, and segmentation predicted from J-Net (Figure 4(a)), U-Net (Figure 4(b)), FCN (Figure 4(c)), and DeepLab-v3+ (Figure 4(d)).

Table 2 shows the results for the EM datasets. We compare our method with FCN and DeepLab-v3+ in EM segmentation tasks. It can be seen from Table 2 that the Dice coefficient of J-Net is 0.9376 and that of other two networks is 0.9185 and 0.9364. On the whole, the performance of J-Net is better than that of the above two algorithms.

The segmentation results of EM datasets by different networks are shown in Figure 5. Figure 5 shows the visual comparison on EM val set. From left to right are input



(a)

(b)

(c)

Figure 5: Segmentation results of EM datasets are compared among our J-Net and other two proposed models. (a) DeepLab-v3+. (b) FCN. (c) J-Net.

images, ground truth, and segmentation outputs from DeepLab-v3+ (Figure 5(a)), FCN (Figure 5(b)), and J-Net (Figure 5(c)).

5. Conclusions

This article analyzed previous works on medical image segmentation, proposed a new architecture (J-Net), and discussed the effect of the information flow path on feature extraction. By connecting ASPP modules in series, changing the encoder network, and reducing the size of the decoder network, an asymmetric encoder-decoder network is designed.

When faced with a complex boundary in segmentation, there is a situation of unstable training (frequent loss fluctuations), mainly because J-Net pays too much attention to high-level semantics and low-level features and neglects to reuse other features. A large number of experiments on the challenging liver datasets and EM datasets have proved the effectiveness of our method. The strategies proposed in this work may be extended to other medical imaging applications and even routine computer vision tasks.

Data Availability

The data used to support the findings of this study are available online.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The work was supported in part by the Science and Technology Innovation Development Program of Jilin City under Grant (no. 201831769 and no. 20190104204) and “Thirteenth

Five-Year" Science and Technology Project of Jilin Province Education Department Contract Number: JJKH20200234KJ.

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communication of the ACM*, vol. 25, pp. 1097–1105, 2012.

[2] J. Redmon, S. Divvala, R. B. Girshick, and A. Farhadi, "You only look once: unified, real-time object detection," in *Proceedings of the 2016 IEEE Conference on Computer Vision And Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016.

[3] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, pp. 1137–1149, 2015.

[4] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 640–651, 2017.

[5] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proceedings of the European Conference on Computer Vision*, pp. 818–833, Zurich, Switzerland, September 2014.

[6] S. Liu and H. Di, "Receptive field block net for accurate and fast object detection," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 385–400, Munich, Germany, September 2018.

[7] O. Ronneberger, P. Fischer, and T. Brox, "U-net: convolutional networks for biomedical image segmentation," in *Proceedings of the International Conference on Medical image computing and computer-assisted intervention*, vol. 9351, Munich, Germany, 2015.

[8] S. Liu, D. Huang, and Y. Wang, "Learning spatial fusion for single-shot object detection," 2019, https://arxiv.org/abs/1911.09516.

[9] D. Zhang, H. Zhang, J. Tang, M. Wang, X. Hua, and Q. Sun, "Feature pyramid transformer," in *Proceedings of the European Conference on Computer Vision*, Glasgow, UK, July 2020.

[10] Q. Chen, Y. Wang, T. Yang, X. Zhang, J. Cheng, and J. Sun, "You only look one-level feature," 2021, https://arxiv.org/abs/2103.09460.

[11] Q. Zhao, T. Sheng, Y. Wang et al., "M2det: a single-shot object detector based on multi-level feature pyramid network," in *Proceedings of the 2019 AAAI Conference on Artificial Intelligence*, vol. 33, Honolulu, HI, USA, 2019.

[12] P. Zhou, B. Ni, C. Geng, J. Hu, and Y. Xu, "Scale-transferrable object detection," in *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 528–537, Salt Lake City, UT, USA, June 2018.

[13] Q. Zhou, Y. Wang, Y. Fan et al., "AGLNet: towards real-time semantic segmentation of self-driving images via attention-guided lightweight network," *Applied Soft Computing*, vol. 96, 2021.

[14] Q. Zhou, J. Wang, J. Liu, S. Li, W. Ou, and X. Jin, "RSANet: towards real-time object detection with residual semantic-guided attention feature pyramid network," *Mobile Networks and Applications*, vol. 26, no. 1, pp. 77–87, 2021.

[15] Z. Liu, Y. Lin, Y. Cao et al., "Swin transformer: hierarchical vision transformer using shifted windows," 2021, https://arxiv.org/abs/2103.14030.

[16] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2018.

[17] L. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, https://arxiv.org/abs/1706.05587.

[18] L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 801–818, Munich, Germany, September 2018.

[19] W. Luo, Y. Li, R. Urtasun, and R. Zemel, "Understanding the effective receptive field in deep convolutional neural networks," 2017, https://arxiv.org/pdf/1910.06041.pdf.

[20] V. Badrinarayanan, A. Kendall, and 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.

[21] T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 106, pp. 936–964, Honolulu, HI, USA, July 2017.

[22] S. Karen and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, https://arxiv.org/abs/1409.1556.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 90, pp. 770–778, Las Vegas, NV, USA, June 2016.

[24] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: scalable and efficient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, 2020.

[25] S. Qiao, L. Chen, and A. Yuille, "Detectors: detecting objects with recursive feature pyramid and switchable atrous convolution," 2020, https://arxiv.org/abs/2006.02334.

[26] G. Ghiasi, T. Lin, R. Pang, and Q. V. Le, "NAS-FPN.: learning scalable feature pyramid architecture for object detection," in *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, 2019.

[27] D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," 2015, https://arxiv.org/abs/1412.6980.

[28] F. Milletari, N. Navab, and S. Ahmadi, "V-net: fully convolutional neural networks for volumetric medical image segmentation," in *Proceedings of the 2016 4th International Conference on 3D Vision (3DV)*, Stanford, CA, USA, 2016.

[29] A. A. Taha and A. Hanbury, "Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool," *BMC Medical Imaging*, vol. 15, no. 1, pp. 29–28, 2015.

[30] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.