Transactions on Computational and Scientific Methods | Vo. 4, No. 12, 2024
ISSN: 2998-8780
https://pspress.org/index.php/tcsm
Pinnacle Science Press

# Advances and Challenges in Deep Learning-Based Medical Image Segmentation

**Amara Khamis[1], Felipe Andrade[2]**
[1]Department of Computer Science, University of Cape Town, Cape Town, South Africa
[2]Department of Computer Science, University of Cape Town, Cape Town, South Africa
*Correspondence should be addressed to Amara Khamis; Amara.khamis555@gmail.com

## Abstract:

The rapid growth of medical image data presents significant challenges for clinicians while offering opportunities to revolutionize diagnostic and treatment practices. Leveraging deep learning technologies, medical image segmentation has emerged as a critical tool for isolating regions of interest (ROI), enabling precise diagnosis and treatment planning. Compared to traditional segmentation methods, deep learning-based approaches eliminate the need for human intervention and deliver superior accuracy and efficiency. However, challenges persist, including limitations in processing high-resolution images, scarcity of annotated medical datasets, and risks of model overfitting due to small-scale training data. Promising research directions include semi-supervised and unsupervised segmentation, as well as the application of generative adversarial networks (GANs) to augment training datasets. These innovations, coupled with advances in hardware and algorithm development, are poised to address existing limitations and significantly enhance medical image segmentation's role in disease diagnosis and management.

## Keywords:

Artificial Intelligence; Medical Image; Diagnosis.

## 1. Introduction

The medical image data has increased by 30% year by year. A large number of medical image data have brought great challenges for clinicians, and also brought opportunities for the reform of diagnosis and treatment modes of diseases. Relying on powerful image recognition, data mining and deep learning technologies, artificial intelligence can effectively solve the problems of medical image big data processing, significantly improve the efficiency and accuracy of data analysis, and enhance the benefits and value of health and diagnosis and treatment [2].Deep learning involves multiple diseases, multiple image modes and multiple omics in intelligent processing and analysis of medical images, and can realize multiple functions. For example, fundus lesion identification, skin cancer identification, brain disease prediction, accurate radiotherapy target delineation for nasopharyngeal cancer, benign and malignant classification of lung nodules, etc. [6].

At present, medical image processing is highly valued at home and abroad. As a key research direction in the process of medical image processing, image segmentation can effectively separate abnormal tissues and structures in the image. It is a condition for reasonable evaluation and appropriate treatment for patients, and it is gradually playing an increasingly important role in the medical field. Image segmentation can extract the specific tissue or structure in the image and provide the quantitative information of the special tissue to the doctor. The images are segmented and can be used in a variety of applications, such as locating diseased tissue, achieving precision injection and clear presentation of tissue structure.

When doctors make diagnosis, only a part of the tissue or structure in the medical image needs to be analyzed. This part of the image is called the Region of Interest (ROI). These ROIs usually correspond to different organs, pathology or some other biological structure. The purpose of medical image segmentation is to segment the ROI in the image and remove the useless information. So far, many medical image segmentation methods have been proposed at home and abroad, and the segmentation methods have experienced the evolution from traditional image segmentation to deep learning-based medical image segmentation.

## 2. Medical Image Segmentation Method based on Deep Learning

### 2.1 FCN

Early deep learning image segmentation algorithms mainly used sliding window method for target segmentation. Sliding window method would produce a large number of redundant candidate regions, and many calculations were repeated, which was inefficient. In addition, the size of image blocks would directly affect the accuracy of segmentation, so it had certain limitations. In 2015, Long et al. proposed FCN, which replaced the traditional sliding window method and was widely used in the field of image segmentation.

The main idea of FCN is to build a network that only includes convolution operations. Input images of any size, and output of the same size can be obtained through effective reasoning and learning. The network structure of FCN is an encoding-decoding network structure model, which replaces the Convolutional layer with the convolutional layer in the classical Convolutional Neural Networks (CNN), so that the whole network is mainly composed of the convolutional layer and the pooling layer, so it is called FCN. In addition, skip connection is designed in the network to connect the global information of the deep network and the local information of the shallow network to compensate each other. In the network structure, the encoder part is mainly used to extract the high-dimensional features of the image, and the spatial dimension of the image decreases after the convolutional layer and pooling layer, while the decoder part is used to upsample the output feature map. The feature map is restored to the same size as the input image, and the extracted high-dimensional features are mapped to each pixel of the final feature map, thus achieving pixel-level image segmentation [3].Compared with the classic CNN network, the advantage of FCN is that it has no limitation on the image size of the input network, but its disadvantage is also not negligible. The pixel-by-pixel classification adopted by FCN ignores the relation between each pixel, does not consider the global context information, and the up-sampling part is an up-sampling operation. If the feature image is expanded 8 times, 16 times and 32 times, the details in the image will be ignored and the result will be fuzzy.

### 2.2 U-net

In order to make full use of high-resolution information for accurate segmentation and the simple and clear semantic information with regular distribution of segmentation targets in human body images, Olaf Ronneberger et al. proposed the U-net network structure, and realized the fusion of semantic information and high-resolution image information through the U-shaped network structure and skip connection. It is suitable for medical image task. U-net network consists of two parts, contraction path and expansion path. Shrink path is used to obtain context information, reduce the spatial dimension of feature graph and increase the number of feature channels. It is divided into four stages. Each stage receives an input and goes through two 3*3 convolution layers, which are activated by modified linear element activation function, and then undersampling is carried out and the maximum pooling operation of 2*2 with step size of 2 is performed. After each stage, the number of characteristic channels doubles. The expansion path is the core of the network, which is used to accurately locate the task target. As with left-hand symmetry, it is divided into four stages, using upsampling to recover target details and spatial dimensions. After each stage the size of the feature map is doubled and the number of features is halved. Finally, through a 1*1 convolution operation, the 64-channel feature graph is converted into a feature graph with a number of categories of 2, and a probability value is output through the sigmoid function, which reflects the possibility of prediction results. The greater the probability, the greater the possibility [4].

### 2.3 Nested U-net

The encoder in the UNet++ network model is the same as the UNet network, which can effectively carry out feature extraction. The feature of the upper layer is sampled to obtain the feature of the next layer, and the feature of the next layer is sampled and fused with the feature of the upper layer, and then fused with the encoder module of the same layer. In the original U-Net structure, the jump connection uses the direct series mode, while the jump connection of U-Net++ uses the dense mode. By adopting the dense connection mode, the network can automatically learn the importance of different depth features during the training process, so that the appropriate number of downsampling layers can be selected according to the needs, and the network parameters can be reduced under the condition of ensuring the network performance. The upper sampling part of the traditional U-Net structure only superimposes the feature map of the lower sampling part of the same layer, and the semantic information of the two layers is quite different, which is not conducive to the optimization of the network [5].U-net ++ + adopts dense connection, and the network can superimpose features from different layers, which reduces the semantic difference between the features of the down-sampling stage and the features of the up-sampling stage, and is more conducive to the optimization of the network. More feature information can also effectively avoid the loss of small target and large target edge information in the original image with the increase of network layers.

### 2.4 Attention Mechanism

In the case of limited computing power, attention mechanism is a resource allocation scheme that is the main means to solve the problem of information overload, allocating computing resources to more important tasks [9].The attention mechanism can strengthen the network's attention to important features, improve the segmentation accuracy of the network on the basis of constant complexity and computation, and is widely used in feature extraction, classification, detection, segmentation and other processing. By explicitly modeling the interdependencies between channels, adaptive recalibration of channel-like feature responses can be achieved by learning to use global information to selectively emphasize information features and suppress redundant information features.

## 3. Technical Difficulties of Medical Image Segmentation

Medical image has some unique characteristics which make the segmentation of medical image more complicated than that of natural image. The specific performance is:

(1) Small amount of data. The scale of finely annotated natural image data is large. Comparatively speaking, it is difficult to obtain medical image data due to complicated annotation and privacy issues. With large amounts of data, the model does not need to be well interpretable, and it is relatively easy to train a good model. When the amount of data is small, sufficient prior knowledge should be provided to the model to ensure that the model can learn key features, and the number of parameters should be controlled to prevent overfitting [7].

(2) Small target. The targets in most medical images are very small, with irregular shapes, blurred boundaries and complex gradients. The segmentation of medical images requires high precision, so more high-resolution information needs to be input to the model to ensure accurate segmentation.

(3) Simple image semantics. The context information of medical images is very important for the diagnosis of human diseases. However, due to the relatively fixed structure of organs, the semantic information in the images is not rich enough. Therefore, the model is required to make full use of the low-resolution information in the training process to ensure the accurate recognition of the target.

(4) multidimensional image. Natural images are all two-dimensional data, while medical images are mostly three-dimensional data. Three-dimensional convolution is needed to extract three-dimensional information in the data, which increases the number of parameters and makes it easy to overfit.

(5) Multi-mode. Compared with natural images, medical images have data of multiple modes, such as OASIS-3 data set, both MRI images and PET images. Data of different modes has its unique characteristics. Models trained on a certain type of data may not be applicable to other data, which requires models to be able to extract the characteristics of different modes, so as to improve the

generalization ability of the model.

These characteristics of medical image determine that encoder - decoder structure network model must be used for medical image segmentation. The high difficulty and complexity of medical image segmentation technology are the main reasons that make medical image segmentation receive special attention in the field of image segmentation.

## 4. Conclusion

Compared with traditional medical image segmentation methods, the segmentation method based on deep learning eliminates human participation and plays an increasingly important role in the field of medical image processing. However, by comparing the literature related to deep learning segmentation, it can be found that there are some difficulties and challenges in the development and evolution of deep learning segmentation network at the present stage.

(1) Nowadays, the resolution of medical images is getting higher and higher, while the current computer hardware equipment can hardly support the processing of high-resolution images, so it is usually necessary to crop the images and send them into the network in blocks for training, which restricts the network to extract more spatial information.

(2) Medical image data set is difficult to obtain. Different tasks in medical image analysis have different requirements on data tagging, and very few data sets are applicable to deep learning models, and medical image data sets are usually small in scale, while the scale of training data directly affects the training effect of deep learning models, too little training data is easy to cause overfitting, resulting in poor performance of the model on other data sets.

Medical image segmentation based on deep learning is of great significance for the diagnosis and treatment of diseases. In order to cope with the above challenges, more and more researchers devote themselves to the field of medical image processing and begin to explore new innovations.

(1) Image segmentation under semi-supervised or unsupervised conditions. It is difficult for the supervised training model to exert its effectiveness against some models which require a large amount of training data. In the absence of annotation data, image segmentation under semi-supervised or unsupervised conditions will be one of the main research directions in the future.

(2) Generative adversarial network generates data sets. Combining image data generated by GAN framework with original data to participate in model training can improve model performance, which is particularly important for medical image analysis. How to divide the original data and generated data reasonably to make the training model achieve the best performance is an important problem that needs to be solved at present and in the future. Image semantic segmentation is widely used. The progress of deep learning in medical imaging attracts experts in computer vision field to solve the task of medical image segmentation. Facing the difficulties in the field of medical image segmentation, the medical image industry is making more efforts to develop new theories and new technologies to open up application prospects. The breakthrough of deep learning in medical image segmentation will make a great contribution to the development of medical field.

## References

[1] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in International Conference on Medical image computing and computer-assisted intervention. Springer, 2015, pp. 234–241.

[2] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, “Unet++: A nested u-net architecture for medical image segmentation,” in Deep learning in medical image analysis and multimodal learning for clinical decision support. Springer, 2018, pp. 3–11.

[3] W. Wang, Y. Li, X. Yan, M. Xiao and M. Gao, "Breast cancer image classification method based on deep transfer learning," Proceedings of the International Conference on Image Processing, Machine Learning and Pattern Recognition, pp. 190-197, 2024.

[4] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz et al., “Attention u-net: Learning where to look for the pancreas,” arXiv preprint arXiv: 1804.03999, 2018.

[5] S. Wang, Z. Liu and B. Peng, "A Self-training Framework for Automated Medical Report Generation," Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 16443-16449, December 2023.

[6] M. Z. Alom, M. Hasan, C. Yakopcic, T. M. Taha, and V. K. Asari, “Recurrent residual convolutional neural network based on u-net (r2unet) for medical image segmentation,” arXiv preprint arXiv: 1802. 06955, 2018.

[7] S. Lu, Z. Liu, T. Liu and W. Zhou, "Scaling-up Medical Vision-and-Language Representation Learning with Federated Learning," Engineering Applications of Artificial Intelligence, vol. 126, Article ID 107037, 2023.

[8] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y.-W. Chen, and J. Wu, “Unet 3+: A full-scale connected unet for medical image segmentation,” in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 1055–1059.

[9] X. Wang, R. Girshick, A. Gupta, and K. He, “Non-local neural networks,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7794–7803.

[10] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, “Ccnet: Criss-cross attention for semantic segmentation,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 603–612.