Proceeding of the IEEE
International Conference on Robotics and Biomimetics
Dali, China, December 2019

An Enhanced U-Net for Brain Tumor Segmentation

Ziru Zhao and Zijian Zhao

School of Control Science and Engineering
Shandong University
Jinan 250061, Shandong, China

***Abstract* - The Convolutional Neural Network (CNN) has been in rapid development these years. It is applied to many fields. Among these, brain tumor segmentation is one of foci. Researchers delves into this with different ways. Some researchers use 3D CNN to handle the brain images for tumor segmentation, and others use 2D CNN modified from models of natural images in semantic segmentation to process separate slices of brain images. Based on properties of brain images and the inherent contradiction of semantic segmentation, we chose 2D U-Net as basic model and modified the skip connection in it to propose our model. To verify effect of our model, we introduced our brain image dataset and made experiments on it and BraTS 2017 dataset. Experiments shows that proposed model outperforms the original U-Net.**

Index Terms - brain tumor segmentation, CNN, U-Net, skip connection.

I. Introduction

The Convolutional Neural Network (CNN) has been in rapid development these years. Its tasks expand from classification [1] to object detection [2], semantic segmentation [3] and many other computer vision tasks. Its input extends from natural images to remote sensing images [4], depth images [5], medical images [6] and other special images. Among these, the semantic segmentation of medical images is one of the foci. The brain tumor segmentation is a hot spot and a challenge.

The biggest difference between brain images and natural images is that one brain image consists of multiple 2D slices to represent a 3D brain, which means brain images contain both 2D information and 3D information. Many researchers take advantage of 3D information for brain tumor segmentation. Kamnitsas et al. [7] used 3D CNN directly, which is a huge consumption of memory. To reduce memory usage, Wang et al. [8] used anisotropic convolution in which they decomposed a 3D kernel with a size of 3×3×3 into a kernel with a size of 3×3×1 and a kernel with a size of 1×1×3. It decreases the amount of parameters but not the size of feature maps. Zhao et al. [9] refined the result of 2D CNN by 3D Conditional Random Fields (CRF). However, CRF is a heavy slowdown to inference and it does not belong to CNN in the strict sense. All of the above models depend on strong 3D information that needs dense 2D slices, which brings a huge cost of image obtainment and label annotation. Besides, human experts always examine 2D slices separately to make diagnosis, which implies 2D information is still valuable to study.

When it comes to 2D information, the brain tumor segmentation is more similar to semantic segmentation of natural images. Many researchers design their structures inspired by the structures of semantic segmentation model. Havaei et al. [10] adopted the early solution of semantic segmentation. They trained a classification model with patches of brain images. However, their inference of whole brain images must be acquired by sliding window which is a tedious operation, and training with patches destroys global information of brain images. Most researchers now focus on end-to-end structures. These structures can be roughly divided into two types: “backbone” and “non-backbone.”

The “backbone” consists of three serial parts: backbone module, feature recognition module and up-sampling module. The backbone module is composed of layers transfer-learned from the pre-trained classification models such as VGG [11], ResNet [12] to get abundant and effective high-level features. The feature recognition module is composed of specially-designed layers to transform features of previous module to semantics. The up-sampling module is composed of one or two up-sampling layers to up-sample semantic feature maps to original resolution and a few other layers to get result. This type of model structure is suitable for natural images [13]. However, researchers hardly adopt this type for brain images, because lack of pre-trained models for brain images make it difficult to implement. Training the backbone module from the ground up is a difficulty considering the huge amount of parameters. Hence, the “non-backbone” is a more common choice.

Without the huge pre-trained models as a good information collector, the “non-backbone” has a smaller but more varying and complicated structure for full use of information. The “non-backbone” can also be divided into two types: “multi-upsample” and “stepwise-upsample.”

The “multi-upsample” is always composed of one down-sampling path and many up-sampling paths. The down-sampling path is composed of serial down-sampling units to get higher-level features gradually. It is similar to backbone module in “backbone” but always smaller and trained from scratch. Every up-sampling path up-samples feature maps from different down-sampling units to original resolution and generates the result. Shen et al. [14] proposed a three-upsampling-paths model with FCN [15] as the down-sampling path for brain tumor segmentation. However, there is a problem in this type. All the up-sampling paths are one-step operations ignoring resolution. Up-sampling low-resolution feature maps to original resolution at single step is an overhasty operation due to lack of spatial details.

The “stepwise-upsample” shows extraordinary talents for this problem. It consists of three parts: down-sampling path,

978-1-7281-6321-5/19//$31.00 ©2019 IEEE

up-sampling path and skip connection. The down-sampling path is composed of serial down-sampling units the same as "multi-upsample." The up-sampling path is composed of serial up-sampling units to up-sample feature maps to original resolution gradually. The above two paths are often symmetrical for skip connection to pass enough information from the former to the latter. The U-Net [16] is a typical and effective case. Many researchers delves into this and modifies the structure. Dong et al. [17] directly applied U-Net to brain tumor segmentation. Shaikh et al. [18] added blocks of densely connected layers into U-Net. Zhou et al. [19] redesigned the skip connection in U-Net, and proposed a nested U-Net architecture.

Inspired by above research, we proposed an enhanced U-Net model, in which skip connection is specially designed. Section II shows reason for our design and structure of model. Section III gives training and test details. Section IV is the conclusion summing up our work.

II. Method

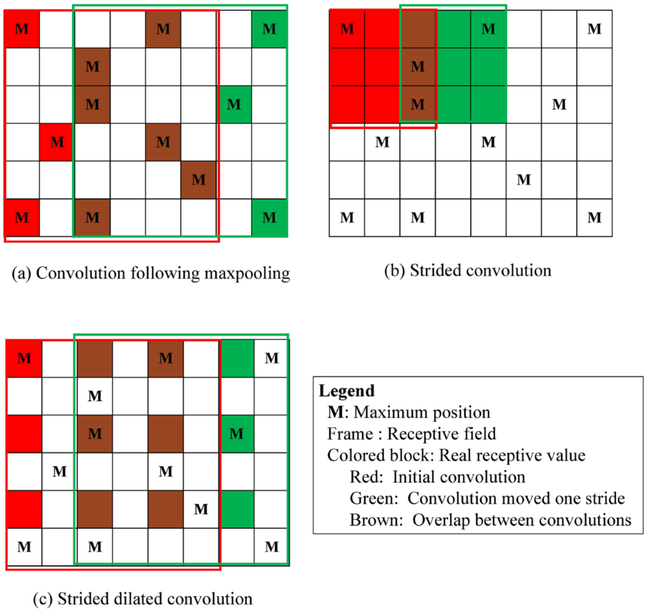
A. *Reason of Design*

There is an inherent contradiction in semantic segmentation. To get semantics, we must pay attention to notable values leaving out details. To get segmentation, we must focus on boundaries that require details. To tackle this contradiction, we redesigned the skip connection of U-Net. We will go through three parts of U-Net to explain the reason of our design in paragraphs below.

The down-sampling layer in down-sampling units is always the max-pooling. It is a very effective operation to get semantics, because max-pooling retains the most notable values discarding others in the area of pooling kernel, which fits semantic characteristics as mentioned in the previous paragraph. Besides, the max-pooling expands receptive field of subsequent convolution, reduces amount of parameters and prevents over-fitting. However, those advantages are obtained in cost of losing resolution and details. The loss of resolution is self-evident. The loss of details can be seen in Fig. 1(a): max-pooling makes subsequent convolution inspect feature maps in a sparse and spatially irregular way. In other word, the spatial invariance of max-pooling results in loss of details.

From above, we can see that the max-pooling is irreplaceable in down-sampling units for semantic obtainment, which leads to inevitable loss of resolution and details, so we turn to up-sampling units for restoring these two. There are three common kinds of up-sampling layers in up-sampling units: bilinear interpolation [20], de-convolution [3] and un-pooling [21]. The bilinear interpolation and de-convolution just restore resolution based on the current information without restoring lost details. The un-pooling though can restore maximum position of corresponding max-pooling, but intermediate layers between the pooling and un-pooling make the corresponding relationship uncertain. In a word, up-sampling units can restore resolution but cannot restore details so far.

The skip connection is the last remaining part for restoring details. The skip connection in original U-Net is one-to-one as shown in Fig. 2(a): the feature maps from the down-sampling unit are directly concatenated with the same-resolution feature maps from corresponding up-sampling unit. However, the low-resolution feature maps from bottom down-sampling unit lose most details, which is worthless for corresponding up-sampling unit to restore details. We should seek for details in upper down-sampling units. This leads to a denser skip connection as shown in Fig. 2(b): All the down-sampling units higher than or equal to an up-sampling unit contribute their feature maps to it. However, high-resolution feature maps cannot be concatenated with low-resolution ones, so we must reduce resolution while preserving details as far as possible. The max-pooling cannot be used as stated before. The strided convolution can reduce resolution without losing details, but it has a smaller receptive field than convolution following max-pooling as shown in Fig. 1(b). The dilated convolution [22] has the same receptive field as max-pooling but cannot reduce resolution. Therefore, the strided dilated convolution (SDC) is a good choice. As shown in Fig. 1(c), SDC inspects feature maps in a sparse but spatially regular (left upper corner) way. This difference from max-pooling makes SDC reduce resolution and preserve more details.


M
Legend
M: Maximum position
Frame : Receptive field
Colored block: Real receptive value
Red: Initial convolution
Green: Convolution moved one stride
Brown: Overlap between convolutions

(a) Convolution following maxpooling

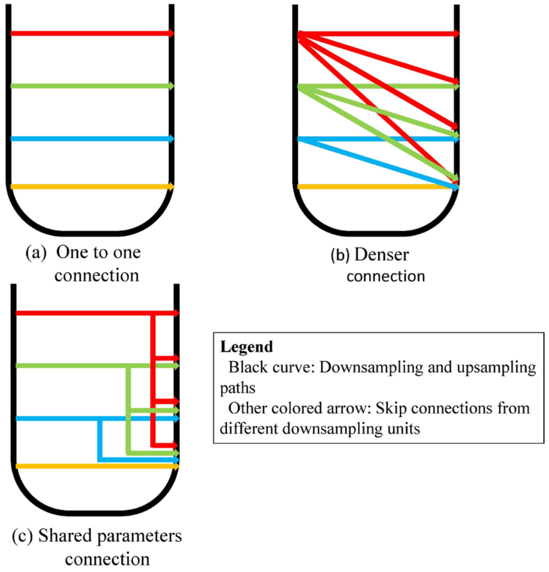
(b) Strided convolution

(c) Strided dilated convolution

Fig. 1 Three ways to reduce resolution.

B. *Detailed Structure of Model*

To reduce amount of parameters, we made our skip connection share parameters, which means that every skip connection from one down-sampling unit converges into a stepwise down-sampling path to several up-sampling units as shown in Fig. 2(c). The detailed structure of our model is as shown in Fig. 3. It is modified from the original U-Net. It has four down-sampling units, each of which is a sequence of two convolutions and one max-pooling, and four up-sampling units, each of which is a sequence of one bilinear interpolation and two convolutions. Every max-pooling and every bilinear interpolation is 2 times down-sampling or up-sampling operation. Convolutions in down-sampling units and two convolutions in bottom are composed of convolution, batch normalization and ReLU, while convolutions in up-sampling units are composed of convolution and ReLU. The final convolution is a 1x1 convolution without ReLU. The hyper parameters of convolutions except the final convolution are: kernel size=3, stride=1, padding=1. The hyper parameters of SDCs are: kernel size=3, stride=2, dilation=2, padding=[2,1,2,1](left, right, top, down).


(a) One to one connection

(b) Denser connection

(c) Shared parameters connection

Legend
Black curve: Downsampling and upsampling paths
Other colored arrow: Skip connections from different downsampling units

Fig. 2 Three kinds of skip connection.

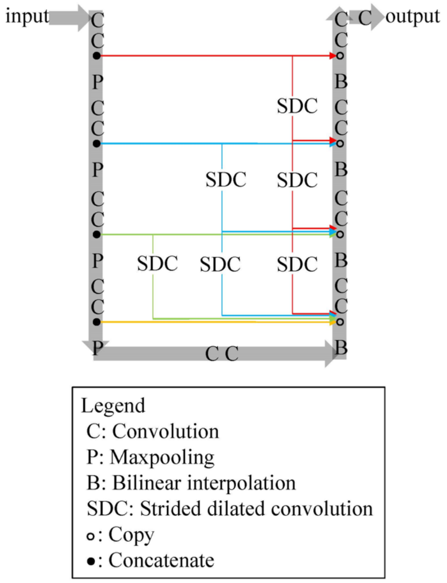


Fig. 3 Structure of our model.

III. Experiments

A. Dataset

1) Our Dataset: Our dataset is composed of 184 transverse multimodal sparse-slices brain images collected from hospitals with permission. Here we give some description about the brain images in dataset and our processing of these.

The sizes of brain images are different. There are sparse 15-50 slices with height and width ranging from 236 to 384 in one brain image. We split slices in brain images into many single slices, and pad slices with zero to 384. The dataset comprises three modalities: T1, T2 and flair. The brain images are not co-registered though they are multimodal, so every modality is independent. Slices are treated as single-channel images without multimodal fusion. We normalized slices to [0,1] through division by maximum to form single-channel inputs to model. The brain images are nearly raw with only annotations by three experts. Annotations comprise tumor (label 1) and non-tumor (label 0).

We used 110 brain images (2233 slices) for training, 37 (754 slices) for validation and the remaining 37 brain images (740 slices) for test.

2) BraTS 2017 Dataset: BraTS 2017 Dataset [23,24,25] is composed of 285 transverse multimodal dense-slices brain images collected from medical institutions. The dataset comprises 210 glioblastoma (GBM/HGG) and 75 lower grade glioma (LGG).We chose HGG because of the ample quantity.

The sizes of brain images are the same. There are dense 155 slices with height and width of 240 in one brain image. We split slices in brain images into many single slices. The dataset comprises four modalities: T1, T1Gd, T2 and FLAIR. The brain images are co-registered to the same anatomical template, interpolated to the same resolution and skull-stripped. Slices can be treated as four-channel images with multimodal fusion. We concatenated slices with different modalities to form four-channel images and normalized images to [0,1] through division by maximum as inputs to model. The brain images are segmented manually by one to four raters, and their annotations were approved by experienced neuro-radiologists. Annotations comprise three parts of tumor (label 1,2,4) and non-tumor (label 0).To keep consistency with our dataset, label 1,2,4 were converted to 1 which means tumor.

We used 126 brain images (19530 slices) for training, 42 (6510 slices) for validation and the remaining 42 brain images (6510 slices) for test.

B. Data Augmentation

We used data augmentation to solve the overfitting. Four methods were used: rotation, flip, resizing and elastic distortion. The parameters are shown in Table I. All values are selected randomly in ranges. The rotation rotates image clockwise (positive value) and anticlockwise (negative value) around the image center. The flip flips image based on vertical medial axis. The resizing resizes image by bilinear interpolation and crop enlarged image to original size. The elastic distortion [26] warps image based on a Gaussian filtered random coordinate displacement matrix. The blank areas after all those transformation are filled with 0.

TABLE I
DATA AUGMENTATION

Methods	Parameters
Rotation	-20° ~20°
Flip	50% probability
Resizing	0.7~1.3 times
Elastic distortion	50% probability, α=720,σ=24

C. Implementation Details

Our model were implemented in PyTorch on two NVIDIA TITAN X GPUs. We used cross entropy as loss function. The cross entropy is shown in (1).

$$cross\ entropy\left(x,Y\right)=-\log\left(\frac{e^{x_Y}}{\sum_{j=1}^{N}e^{X_J}}\right) \qquad (1)$$

where x represents pixel, Y represents the true class of x, N represents the number of classes, j represents any classes, x_y and x_j represents the output probability of pixel x belonging to Y or j. In our experiment, N=2, and Y means tumor.

We used Adaptive Moment Estimation (Adam) for training, with initial learning rate 1x10^{-3}, weight decay 5x10^{-7}, batch size 32, and maximal iteration 500.

D. Results

We used Dice, Sensitivity, and Specificity for evaluation. The three metrics are shown in (2) (3) (4).

$$Dice=\frac{2\left|P_1\cap T_1\right|}{\left|P_1\right|+\left|T_1\right|} \qquad (2)$$

$$Sensitivity=\frac{\left|P_1\cap T_1\right|}{\left|T_1\right|} \qquad (3)$$

$$Specificity=\frac{\left|P_0\cap T_0\right|}{\left|T_0\right|} \qquad (4)$$

P_1 is predicted region of tumor, and T_1 is true region of tumor. P_0 is predicted region of non-tumor, and T_0 is true region of non-tumor.

We evaluated segmentation results of test sets on our enhanced U-Net, U-Net, FRRN [27] and Attention U-Net [28]. The test results can be seen in Table II and Table III.

TABLE II
TEST RESULTS ON OUR DATASET

	Dice	Sensitivity	Specificity
Enhanced U-Net	**0.7948**	**0.7152**	**0.9986**
U-Net	0.7793	0.7080	**0.9986**
FRRN	0.7847	0.7129	0.9985
Attention U-Net	0.7576	0.7044	0.9976

TABLE III
TEST RESULTS ON BRATS 2017

	Dice	Sensitivity	Specificity
Enhanced U-Net	**0.8911**	**0.8592**	0.9981
U-Net	0.8862	0.8544	0.9980
FRRN	0.8851	0.8531	0.9980
Attention U-Net	0.8900	0.8530	**0.9983**

From the Table II and Table III, we can see that our model outperformed other models in most metrics. Fig. 4 shows segmentation results from one brain slice in our dataset. Fig. 5 shows results from one slice in BraTS 2017. Fig. 6. shows the magnified difference between each output and ground truth of the slice from BraTS 2017 for a clear comparison.

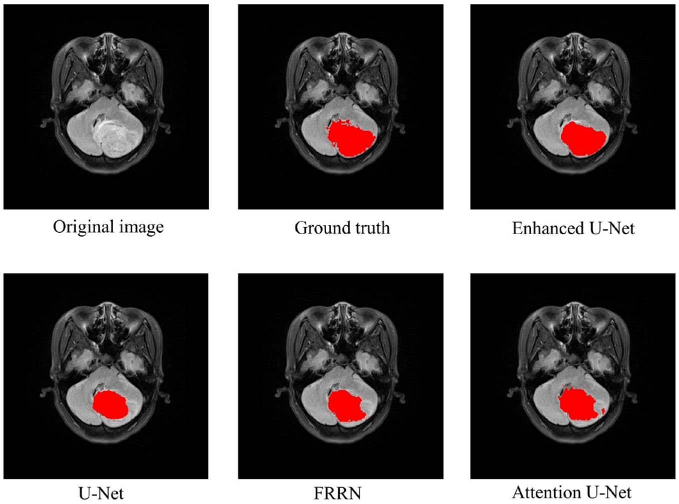

Original image Ground truth Enhanced U-Net
U-Net FRRN Attention U-Net

Fig. 4 Segmentation results of one slice from our dataset

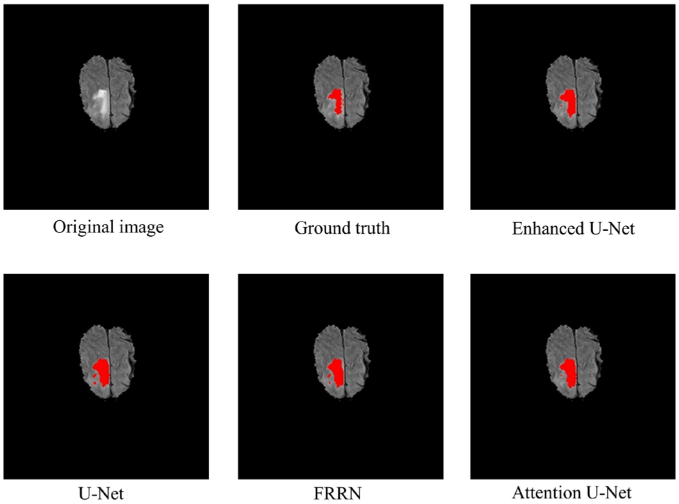

Original image Ground truth Enhanced U-Net
U-Net FRRN Attention U-Net

Fig.5 Segmentation results of one slice from BraTS 2017

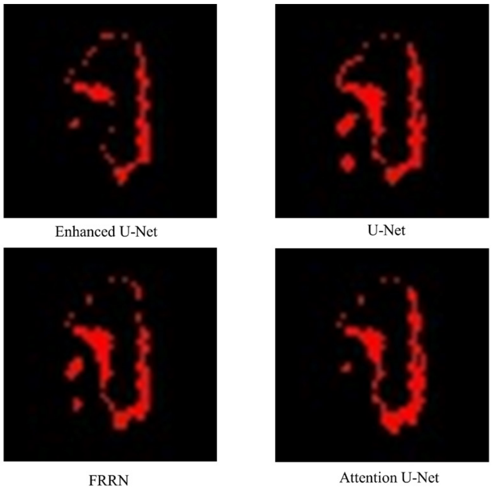

Enhanced U-Net U-Net
FRRN Attention U-Net

Fig.6 Differences from BraTS 2017

Images in our dataset are harder for models to distinct because of little pre-processing. We can see that most models except enhanced U-Net missed a big part of tumor because of loss of details. Enhanced U-Net can utilize more details to make better predictions in hard cases.

Images in BraTS 2017 are standard and neat due to fine pre-processing. All the models got similar results. The difference of enhanced U-Net is less than others as shown in Fig. 6. That implied enhanced U-Net can still use details to acquire more accurate boundary even in easy cases. All of above proved that our design truly preserved and utilized more details for prediction in easy or hard cases.

IV. Conclusion

In section I, we introduced research status about brain tumor segmentation and related models. We chose U-Net as base model according to the introduction. In section II, we pointed out the contradiction in semantic segmentation, and tried to solve this by making a denser skip connection in which we use SDC to reduce resolution with preserving more details. In section III, we described datasets and experiments configuration. Experiments show that our model outperform other models. Besides, the idea of denser skip connection and reducing resolution with SDC can be easily applied to other tasks that need details. However, there is still possibility to improve performance of our model, in part because SDC still lose some details. In the future work, we decide to combine our idea with a more complicated model for better performance. We also plan to find out a way to preserve more details than SDC.

Acknowledgment

This work has been supported by the Specialized Research Fund for the Doctoral Program of Higher Education of China (No.20130131120036), the Promotive Research Fund for Excellent Young and Middle-aged Scientists of Shandong Province (No.BS2013DX027), the National Natural Science Foundation of China (No.81401543, 61673244), the French National Research Agency (ANR) through TecSan Program. We give our sincere gratitude to above supports.

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” Advances in neural information processing systems, pp. 1097-1105,2012.

[2] S. Ren, K. He, R. Girshick and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” Advances in neural information processing systems, pp. 91-99, 2015.

[3] H. Noh, S. Hong and B. Han. “Learning deconvolution network for semantic segmentation,” Proceedings of the IEEE international conference on computer vision, pp. 1520-1528, 2015.

[4] E. Maggiori, Y. Tarabalka, G. Charpiat and P. Alliez, “Convolutional neural networks for large-scale remote-sensing image classification,” IEEE Transactions on Geoscience and Remote Sensing, vol. 55, no. 2, pp. 645-657, 2017

[5] L. Ge, H. Liang, J. Yuan and D. Thalmann, “3d convolutional neural networks for efficient and robust hand pose estimation from single depth images,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1991-2000, 2017.

[6] N. Tajbakhsh, et al, “Convolutional neural networks for medical image analysis: Full training or fine tuning?,” IEEE transactions on medical imaging, vol. 35, no. 5, pp. 1299-1312, 2016

[7] K. Kamnitsas, et al, “Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation,” Medical image analysis, vol. 36, pp. 61-78, 2017.

[8] G. Wang, W. Li, S. Ourselin and T. Vercauteren. “Automatic brain tumor segmentation using cascaded anisotropic convolutional neural networks,” International MICCAI Brainlesion Workshop, pp. 178-190, Springer, Cham, 2017

[9] X. Zhao, Y. Wu, G. Song, Z. Li, Y. Zhang, and Y. Fan. “3D brain tumor segmentation through integrating multiple 2D FCNNs.” In International MICCAI Brainlesion Workshop, pp. 191-203. Springer, Cham, 2017.

[10]M. Havaei, et al, "Brain tumor segmentation with deep neural networks," Medical image analysis, vol.35 pp. 18-31, 2017.

[11]K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv preprint, arXiv:1409.1556, 2014.

[12]K. He, X. Zhang, S. Ren and J. Sun, “Deep residual learning for image recognition,” Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778, 2016.

[13]L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff and H. Adam, “Encoder-decoder with atrous separable convolution for semantic image segmentation,” Proceedings of the European Conference on Computer Vision (ECCV), pp. 801-818, 2018.

[14]H. Shen, J. Zhang and W. Zheng, “Efficient symmetry-driven fully convolutional network for multimodal brain tumor segmentation,” 2017 IEEE International Conference on Image Processing (ICIP), pp. 3864-3868, 2017.

[15]J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3431-3440, 2015.

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

[17]Hao Dong, Guang Yang, Fangde Liu, Yuanhan Mo and Yike Guo. “Automatic brain tumor detection and segmentation using U-Net based fully convolutional networks,” Annual conference on medical image understanding and analysis, pp. 506-517, Springer, Cham, 2017.

[18]M. Shaikh, et al, “Brain tumor segmentation using dense fully convolutional neural network,” International MICCAI Brainlesion Workshop, pp. 309-319, Springer, Cham, 2017.

[19]Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang. “Unet++: A nested u-net architecture for medical image segmentation,” Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, pp. 3-11. Springer, Cham, 2018.

[20]M. Jaderberg, K. Simonyan and A. Zisserman, “Spatial transformer networks,” Advances in neural information processing systems, pp. 2017-2025, 2015.

[21]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.

[22]F. Yu and V. Koltun, “Multi-scale context aggregation by dilated convolutions,” arXiv preprint, arXiv:1511.07122, 2015.

[23]B. H. Menze,, et al, “The multimodal brain tumor image segmentation benchmark (BRATS),” IEEE transactions on medical imaging, vol. 34, no. 10, pp. 1993-2024, 2014

[24]S. Bakas, et al, “Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features,” Scientific data, 4: 170117,2017

[25]S. Bakas, et al, “Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge”, arXiv preprint, arXiv:1811.02629, 2018

[26]P. Y. Simard, D. Steinkraus, and J. C. Platt. “Best practices for convolutional neural networks applied to visual document analysis.” Icdar. Vol. 3. No. 2003, 2003.

[27]T. Pohlen, A. Hermans, M. Mathias and B.Leibe, “Full-resolution residual networks for semantic segmentation in street scenes,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4151-4160, 2017.

[28]O. Oktay, et al, “Attention u-net: Learning where to look for the pancreas”, arXiv preprint arXiv:1804.03999, 2018.