A Triple Residual Multiscale Fully Convolutional Network Model for Multimodal Infant Brain MRI Segmentation

Yunjie Chen[1*], Yuhang Qin[1], Zilong Jin[2], Zhiyong Fan[3] and Mao Cai[1]
[1] School of math and statistics, Nanjing University of Information Science & Technology, Nanjing 210044, China
[e-mail: priestcyj@nuist.edu.cn, qinyuhang98@163.com, 1440896206@qq.com]
[2] School of computer and software, Nanjing University of Information Science & Technology, Nanjing 210044, China
[e-mail: zljin@nuist.edu.cn]
[3]School of Automation, Nanjing University of Information Science & Technology, Nanjing 210044, China
[e-mail: 001163@nuist.edu.cn]
*Corresponding author: Yunjie Chen

Received January 6, 2020; revised February 25, 2020; accepted March 20, 2020; published March 31, 2020

Abstract

The accurate segmentation of infant brain MR image into white matter (WM), gray matter (GM), and cerebrospinal fluid (CSF) is very important for early studying of brain growing patterns and morphological changes in neurodevelopmental disorders. Because of inherent myelination and maturation process, the WM and GM of babies (between 6 and 9 months of age) exhibit similar intensity levels in both T1-weighted (T1w) and T2-weighted (T2w) MR images in the isointense phase, which makes brain tissue segmentation very difficult. We propose a deep network architecture based on U-Net, called Triple Residual Multiscale Fully Convolutional Network (TRMFCN), whose structure exists three gates of input and inserts two blocks: residual multiscale block and concatenate block. We solved some difficulties and completed the segmentation task with the model. Our model outperforms the U-Net and some cutting-edge deep networks based on U-Net in evaluation of WM, GM and CSF. The data set we used for training and testing comes from iSeg-2017 challenge (http://iseg2017.web.unc.edu).

Keywords: Isointense phase, Tissue segmentation, Convolutional network, Residual multiscale block

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions to improve the quality of the paper. This work was supported in part by the National Nature Science Foundation of China 61672291, 61976241, 61602252, Six Talent Climax Foundation of Jiangsu Province of China SWYY-034, the Natural Science Foundation of Jiangsu Province of China BK20191394, BK20160967, and the Key Research Project of Jiangsu Province of China 2019A005.

http://doi.org/10.3837/tiis.2020.03.003 **ISSN : 1976-7277**

1. Introduction

Image segmentation is an important preprocessing of image recognition and computer vision. During the development of medical imaging technology, image segmentation is of great significance in medical applications. So far, many methods are applied to medical images analysis, including some traditional methods based on statistical analysis and partial differential equations. Furthermore, with the appearance of CNN, many segmentation methods based on deep learning are increasingly proposed and gradually replaced the traditional algorithms into the mainstream. The emergence of U-Net opened up the regulation of convolutional neural network to segment biomedical images.

U-Net is a segmentation network proposed by Ronneberger et al., "in the ISBI Challenge [1] which is inspired by the FCN [2]. **Fig. 1** shows the overall structure. The neural network is mainly composed of two components: a contracted path and an extended path. The contracted path is mainly used to capture the context information of the image. The input via two 3×3 convolution layers and a maxpooling layer, whose process is repeated four times and feature maps are reduced to 1⁄64. The extended path precisely locates the part to be segmented in the image. The result of the contracted path via a deconvolution layer to expand the size and two 3×3 convolution layers, whose process is repeated four times to make feature maps restored to the original size. Skip connection is also used in the network to transfer the shallow feature maps to the upper layer symmetrically, making better use of the information of different scales. In many cases, the training of deep learning networks requires a large number of data sets, and the cost of biomedical data (images and texts) is higher. The U-Net is very effective for the segmentation of medical images with few samples and also has good noise immunity. To a certain extent, the noise image has less influence on training process. However, this model also has problems objectively: 1) most medical images have weak edges, which make the network perform better classification difficultly and cause partial loss of details; 2) structurally, simply superimposing the convolution layer can improve the expression ability of network, which will increase a mass of parameters and make training network difficult. Up to now, many scholars have proposed many improved methods for the U-Net [3, 4, 10, 12, 13].

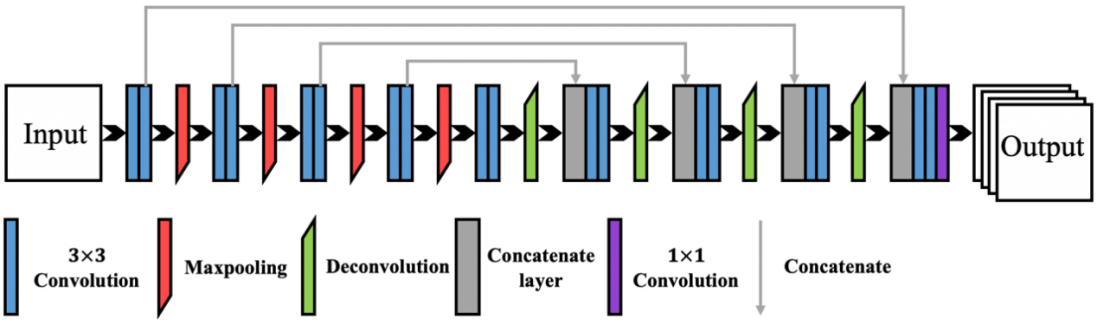

Fig. 1. The U-Net structure includes convolution, maxpooling, concatenate and deconvolution.

Chen L et al., "proposed DRINet [3] and a fully automatic acute ischemic lesion segmentation model (EDD+MUSCLE Net) [4]. The DRINet absorbed the idea of DenseNet [5], ResNet [6] and Inception V1 [7] and adopts dense connection block, residual inception block and unpooling block. Compared to U-Net, the structure of DRINet does not have any

skip connections and the connections are packaged in blocks, which makes the method be more flexible and has more accurate segmentation results. The EDD+MUSCLE Net [4] combines EDD Net with MUSCLE Net. The EDD Net consists two parallel full convolution network architectures, which can obtain the integrated segmentation results. MUSCLE Net composed of a mini VGG-Net [8, 9] is applied to judge true or false positive accurately. The fully automatic acute ischemic lesion segmentation model has eminent ability of segmentation and recognition using for tumor images. Some details are lost in the segmentation results because of the simple network structure and fewer network layers.

Foivos I et al., "proposed ResUNet-a [10] which introduced a residual block to eliminate the problem of gradient dispersion or explosion effectively. The addition of the pyramid scene parsing pooling layer uses background context information [11], which strengthens the use of the whole network feature information and improves the performance of network. Concurrently, they improved Dice loss function to expedite the convergence of the network. However, the performance in segmentation results is not ideal in details.

MDU-Net was proposed by Zhang J et al., "[12] and created three multi-scale dense connected structure: Dense Encoder and Decoder Block, Dense Cross connections Block and Fully Multi-scale Dense connected U-shape architecture. The characteristic of dense connection is lifting gradient back propagation, which is fully applied in the U-shape structure to make the training easier. Zhou Z et al., "created a nested U-Net architecture [13] that integrates different levels of feature information. The application of the deep supervision at the end of structure contributes to update weights more quickly during training. The skip connections are various in the two methods, which also brings a lot of calculation in the training process due to the large amount of parameters

Badrinarayanan V proposed the SegNet method [14] which is extremely similar to FCN. The pooled index in upsampling layers make the model easier to optimize. The PSPNet [15] creates a pyramid pool block that can aggregate context information of different regions, which enhances the utilization of global information and enormously improves the performance of scene parsing. Nevertheless, these two methods have achieved good effect in scene image segmentation, but poor consequence in medical image segmentation.

DeepLab [16] applied the atrous convolution which expands the receptive field, but the calculation amount is the same as convolution. The receptive field can extract more global information in images. They improved the performance of image boundary positioning by introducing a full connected Condition Random Field (CRF) in the final output of network. Although the model is successful, they still exist at least two limitations. Firstly, it needs to perform convolutions on a large number of detailed feature maps that usually have high-dimensional features, which compute expensively . Moreover, a large number of high-dimensional and high-resolution feature maps also require huge amounts of GPU memory resources [17].

2. Related Work

The brain diseases are major killers of human health and are difficult to cure. Accurate diagnosis and efficient treatment of brain diseases has always been a medical problem. In the first year of human birth, the brain is in the stage of rapid development, some neuropathic diseases, such as attention deficit hyperactivity disorder (ADHD), infantile autism [18], bipolar affective disorder and schizophrenia, may be reflected in the pathological brain tissue of the patient's infancy. Therefore, it is significant that brain tissues are accurately segmented in infant images.

T1-weighted (T1-w) and T2-weighted (T2-w) non-invasive infant brain multimodal magnetic resonance images (MRI) are common, which provide enough data and good conditions for our segmentation studies. Our essential research process is to accurately segment three types of brain tissue (CSF, GM and WM), which is also very important for registration [19] and atlas building [20, 21]. However, the T1-weighted and T2-weighted infant brain MR images have problems of low contrast and uneven gray scale. **Fig. 2** demonstrates the gray distribution diagram of tissues in the MR images, from which it can be observed that the gray values in three tissues have a high overlap. In the gray distribution diagram of T1-w image, the gray values are largely overlapping in GM and WM. Moreover, all tissues are largely overlapping in the gray distribution diagram of T2-w image. The situation indicates that the contract is the lowest, which is the challenge for us to finish the tissue segmentation task. Obviously, some traditional methods are disabled to solve such problems, which leads us to attempt to use deep learning to complete the challenges.

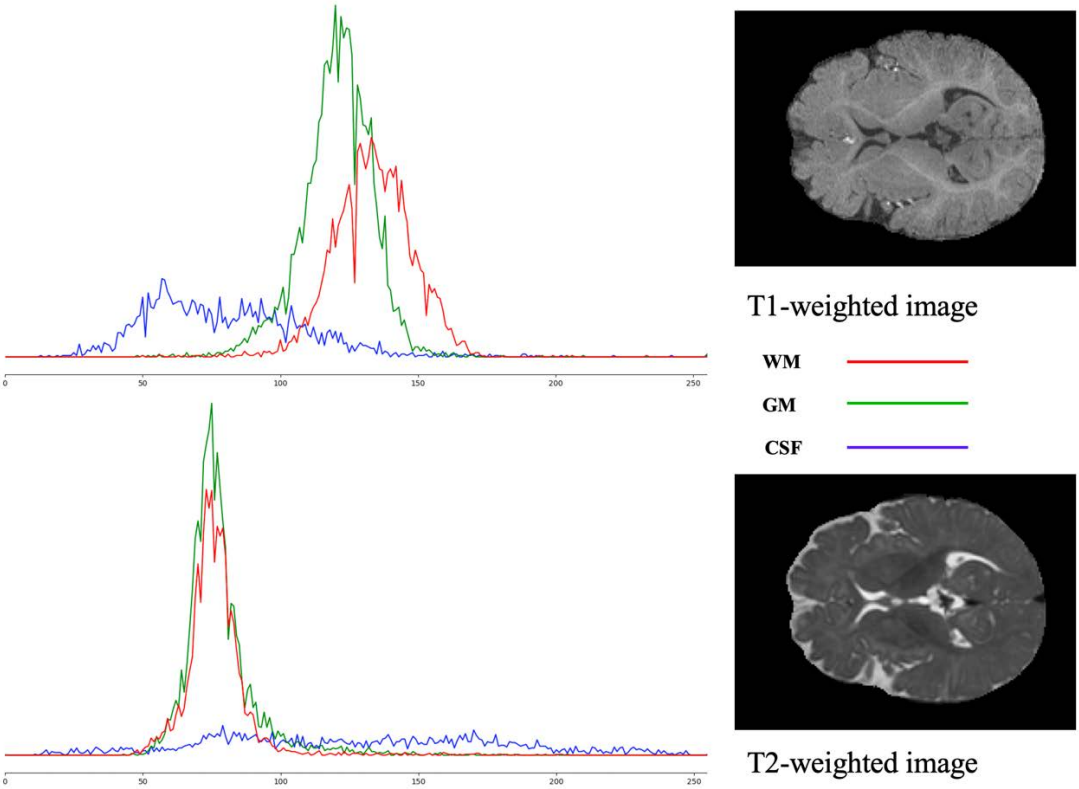

Fig. 2. The T1- and T2-weighted infant MR images in the isointense phase exhibit the contrast and obvious difference. The left side of the MR images are the tissue distribution in the brain, among them, the distribution overlap of WM and GM is very high.

In structures of deep learning network, a large quantity of layers is superimposed, which means that many more features can be extracted to make the network have stronger expressive ability. However, increasing the number of layers will also bring plenty of parameters to make training more difficult，which will make convergence becoming slower and bring about lower accuracy (as opposed to shallower networks). The residual learning can effectively solve such a problem [6]. In each bottleneck block, the output is the addition of the convolution result and an identity map. The skip connection makes the convolution kernel only Learning the residual

features between the input and the output, which makes the gradient not disappear during the transfer process. Therefore, the network training becomes easy. The Inception v1[7] is proposed to break the conventional convolution pattern. The creation of the inception module reduced a good deal of parameters and saved computational overhead. At the same time, the design of this module can effectively expand the scope of expression features.

U-Net and its variant models have the defects of losing details in many data sets of image segmentation task, and some of them even have problems such as too long training time, gradient disappearance and slow convergence. In order to solve these problems, we propose a triple residual multiscale fully convolution network (TRMFCN) model with three levels of input, which can extract from multiple scales effectively. Moreover, we introduce the Residual Multiscale (RM) block to make the convergence more easily and apply Concatenate Block to extract more information. Our main contributions are: 1) the multi-scale input method increases the utilization of image information; 2) the Residual Multiscale (RM) block structure improves the computational efficiency; 3) The sufficient concatenations between layers enhance the reusability of feature information.

3. TRMFCN

3.1 Triple-Branched model

Fig. 3 depicts the overall structure of our proposed model (TRMFCN), which is composed of encode and decode process. Encode process includes: traditional 2d convolution, maxpooling and residual multiscale (RM) block. Decode process includes: deconvolution, residuals multiscale (RM) block, concatenate block and traditional 2d convolution. The RM block is inspired by ResNet [6] and Inception V1 [7]. The creation of this concatenate block is inherited from U-Net.

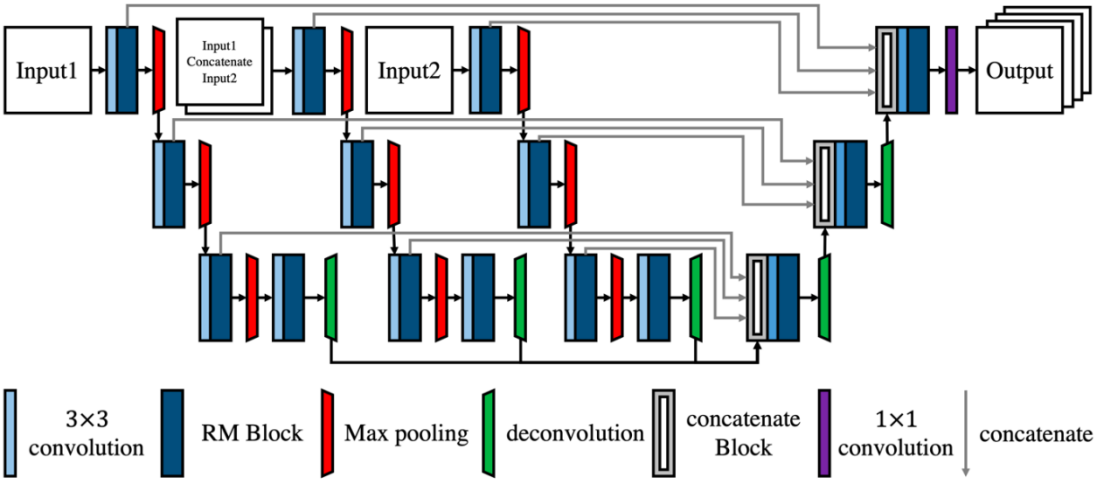

Fig. 3. The entire process is shown in the TRMFCN structure map, RM Block and Concatenate Block are important parts of the structure, the details of them are described in Fig. 4 and Fig. 5, respectively.

In TRMFCN structure, the 2d convolution with 3×3 kernel-size is placed before all RM blocks to increase feature maps, and the kernel number was continuously increased to 60, 120, 240 and 480 in the encoding process, continuously decreased to 240, 120, and 60 in the decoding process. To enable the RM block to extract enough features, we have placed four 2d

convolutions with a kernel size of 3×3 in front of the model output. Moreover, we also placed four 2d convolutions with a kernel size of 1×1 at the end of the model to obtain desired classification results (four classes). We set all the deconvolution step sizes to 2 to restore the image step by step. After 3 times of upsampling, the feature maps are restored to the original image size. The advantages including: 1) the form of input strengthens the information extraction of the data; 2) input branches can be adjusted according to different data sets, and the structure is flexible.

3.2 Residual Multiscale block

In the feature extraction process, a large-scale convolution kernel can make the model learn some large features, and a small-scale convolution kernel can make the model learn some details. **Fig. 4** is the residual multiscale (RM) block, the basic convolution is divided into three different convolutions, and the number of convolution kernels is one third of the input channels. Finally, we put the concatenation of all convolution results and the input into the shortcut (addition). The essence of shortcut connection is the identity map. our block output can be expressed as follows:

$$x_{k+1} = g\left(\left(f_1(x_k) \odot f_3(x_k) \odot f_5(x_k)\right), x_k\right) \tag{1}$$

where the $\odot$ indicates concatenation and $g(\cdot)$ denotes the identity mapping. $f_1(\cdot)$, $f_3(\cdot)$ and $f_5(\cdot)$ represent the convolution of 1×1, 3×3 and 5×5 kernel sizes after batch normalization (BN) [22] and rectified linear unit (ReLU) [23], respectively.

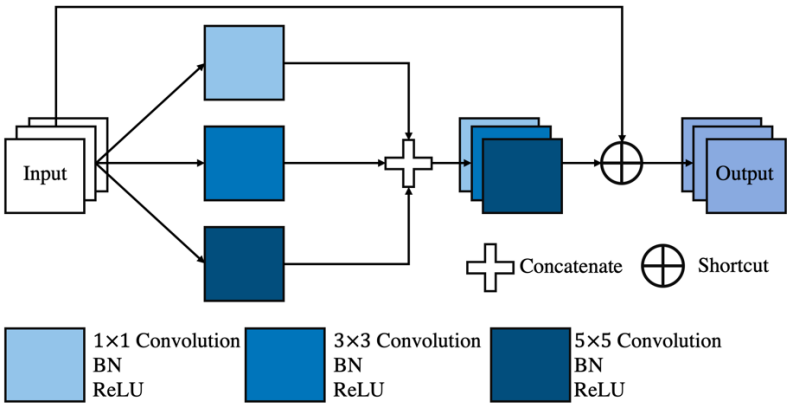

Fig. 4. A residual multiscale block is the inception module combines residual connection, different branches have convolution kernels of different sizes, a skip connection appears after feature maps concatenated.

Unlike the traditional 2d convolution, our block can extract more information from feature maps. The residual connections [24] make the training easier, and the multiscale convolutions make the model learn more features.

3.3 Concatenate Block

Fig. 5 demonstrates the structure of concatenate block, from which we concatenate feature maps from the encoding process and decoding process (the size of the fusion here is the same) to complete feature fusion. Moreover, fusion elements in encoding process have two more parts than U-Net, and **Fig. 1** and **Fig. 3** clearly show the difference. The concatenate layer

fuses many more shallow features, which improve the reusability of features and contribute to the decoding process.

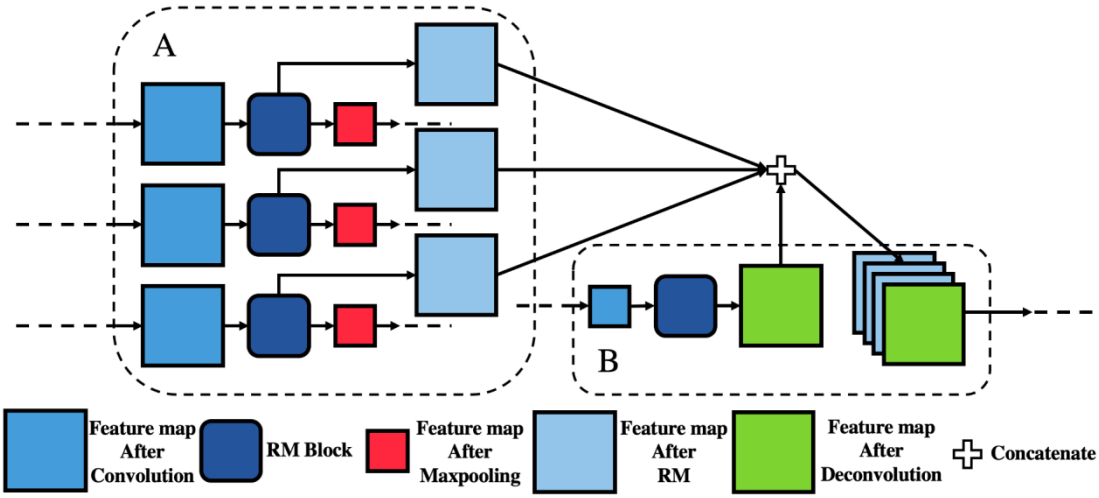

Fig. 5. A is the encoding part of TRMFCN, B is its decoding part, and the Concatenate block represents the way of feature fusion.

3.4 Fully convolutional networks

Convolutional neural network (CNN) usually connects several full connection layers after a series of convolution, and finally obtains the final feature corresponding to the image, namely probability vector, which is used for image level classification or regression task. FCN [2] performs the pixel-level classification task and turns the full connection into convolution and deconvolution, then the feature maps are restored to the original image size. The specific formula can be expressed as follows:

$$C_i = W_i \otimes C_{i-1} + b_i \tag{2}$$

$$C_i = \sigma(C_i) \tag{3}$$

where C_i denotes the result of the i-th convolution, W_i denotes the i-th convolution weight vector and b_i denotes the i-th convolution bias, $\otimes$ is the convolution, and the $\sigma(\cdot)$ is the activation function. The objective function, namely the energy function, is defined as:

$$E = -\sum_{I_{ij} \in I} \sum_{c=1}^{C} q_c(I_{ij}) \log p_c(I_{ij}) \tag{4}$$

The $p_c(I_{ij})$ represents the output probability, which is pixel I_{ij} belongs to class c, and $q_c(I_{ij})$ is the true probability distribution.

3.5 Fully convolutional networks

In order to reasonably analyze the segmentation result of different models, we use Jaccard similarity (Js value) [25] to evaluate the performance of different models, which is defined as:

$$Js(S_1, S_2) = (S_1 \cap S_2)/(S_1 \cup S_2) \tag{5}$$

Where S_1 denotes the segmentation result and S_2 denotes the ground truth, in this way, a higher JS value indicates that the model performs better in the test. In the subsequent evaluation, Js values of CSF, GM and WM will be taken as important indicators to evaluate the model quality. The expressions of the three parts can be expressed as follows:

$$Js_{CSF}(S_{CSF1}, S_{CSF2}) = (S_{CSF1} \cap S_{CSF2})/(S_{CSF1} \cup S_{CSF2}) \tag{6}$$

$$Js_{GM}(S_{GM1}, S_{GM2}) = (S_{GM1} \cap S_{GM2})/(S_{GM1} \cup S_{GM2}) \tag{7}$$

$$Js_{WM}(S_{WM1}, S_{WM2}) = (S_{WM1} \cap S_{WM2})/(S_{WM1} \cup S_{WM2}) \quad (8)$$

Where S_{CSF1} denotes the segmentation result of cerebrospinal fluid (CSF), and S_{CSF2} denotes the ground truth of CSF. The S_{GM1} denotes the segmentation result of gray matter (GM), and the S_{GM2} denotes the ground truth of GM. The S_{WM1} denotes the segmentation result of white matter (WM), and the S_{WM2} denotes the ground truth of WM.

4. Experiments

4.1 Segmentation in infant brain MR images

Data preprocessing: The dataset we choose for the experiment is from iSeg-2017 challenge, and the average age of these babies collected was 6 months [26] without any pathology. We selected 10 labeled babies, each with 256 groups (256 brain MR T1-w images, 256 brain MR T2-w images and corresponding labeled images). Since two-thirds of the images for each baby are all black backgrounds (all pixel values are 0), we removed these images to avoid interfering with network training. At last, a total of 996 images were obtained and the data was scrambled. About 10% (96 groups) were selected as the test set and the remaining 10% as the verification set.

Model train: The encoding structure of our model (TRMFCN) has three inputs: the first input is the T1-w image with the size of $192 \times 144 \times 1$; the second input is the T2-w image, and its size is also $192 \times 144 \times 1$; the third input is different from the first two, which concatenates T1-w and T2-w images with a size of $192 \times 144 \times 2$. There was only one input in the experiments of comparison methods. Therefore, we concatenate T1-w image and the corresponding T2-w image with a size of $192 \times 144 \times 2$ as the input.

The Adam optimizer is chosen in the training process of our model. We set mini-batch size to 10 and the iterations to 80. We also apply the dropout [27] to prevent overfitting. After getting the result of the residual multiscale block, we set the dropout rate to 0.3. Our experiment used the Keras package in python codes. The training lasts for 2 hours, and it takes an average of 90 seconds to complete an iteration. The entire training process was on an NVIDIA GeForce GTX 1080 Ti GPU (11GB).

Experiment Results: We used the seven methods: FCN, U-Net, DRINet, EDD+MUSCLE Net, MDU-Net, ResUNet-a and U-Net++ as comparative experiments. Each method was trained 8 times and the best model was saved, and we randomly selected one for test and comparison. We use these trained models to test 96 T1-w images and the corresponding 96 T2-w images. The Js values (CSF, GM and WM) of the test results are listed in **Table 1** clearly.

From the 7 comparison results, we can intuitively see the differences, Firstly, the Js values (CSF, GM and WM) of U-Net in all classes are far superior to FCN, which indicates the positive function of concatenation in segmentation. Although DRINet constitutes a simple network structure in a modular way and removes the skip connection, the performance is not as good as the original U-Net in this data set. Additionally, the other three methods: EDDNet, MDUNet and U-Net++ are also worse than U-Net. Moreover, the ResUNet-a outperformed U-Net on WM and GM, which indicates that the residual connection optimizes the model to some extent.

Compared to the results of these seven methods, the model we proposed demonstrates the best results, whose test results of the model obtained the highest Js values (CSF, GM and WM) in eight methods. We have concluded two main reasons that improve the accuracy: 1) Residual

learning makes our model converge faster. 2) Various concatenations allow the network to learn more features from shallow layers.

We randomly select one image from 96 test result images to visually compare the segmentation results of different methods. **Fig. 6** (a-g) shows the segmentation results of seven methods, where the **Fig. 6** (a-f) shows the images of six comparison methods and the **Fig. 6** (g) is the result of our method. The **Fig. 6** (h) is the corresponding label. Due to the poor segmentation results of the FCN, we did not show the renderings. We marked the label and all segmentation images with a dark red box in the same representative position. Compared the differences in these red boxes, we find that white matter bands and cerebrospinal fluid bands have evident fractures in the segmentation images of the six comparative methods. These methods lost some details in the brain tissue. Obviously, these methods do not perform as well as our method in segmentation of baby brain MR images.

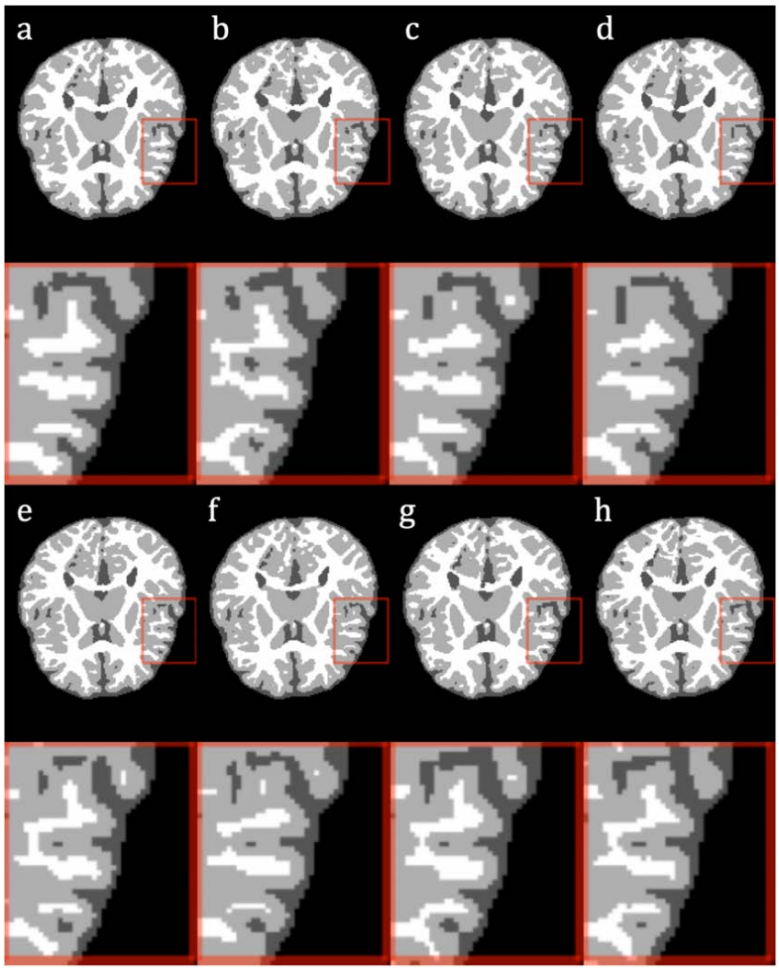

Fig. 6. Comparison of seven segmentation methods. From (a) to (h): U-Net, DRINet, EDDNet, MDUNet, ResUNet-a, U-Net++, the results of the proposed method, and the ground truth. The dark red box represents the obvious segmentation error.

Table 1. Comparison of mean Js values of CSF, GM, WM% in eight methods

Method	CSF (%)	GM (%)	WM (%)
FCN [2]	52.35	62.17	51.21
U-Net [1]	87.96	80.70	72.24
DRINet [3]	81.61	78.75	69.81
EDDNet [4]	85.86	78.92	70.28
MDUNet [12]	86.65	78.05	67.56
ResUNet-a [10]	87.71	81.50	74.34
U-Net++ [13]	84.55	76.54	65.24
TRMFCN (our)	89.26	83.84	77.26

We obtained the accuracy (Acc) and variance (Var) of the 350 test results by calculation in **Table 2.** The accuracy (Acc) can be expressed as follows:

$$Acc(S_1', S_2') = (S_1' \cap S_2')/S_2' \tag{9}$$

where the S_1' denotes the segmentation result without black background information. The S_2' denotes the corresponding ground truth without black background information. Our model obtained the lowest variance in accuracy (Acc), CSF and GM, which manifests that our model has good robustness.

Table 2. Methods of accuracy (%) and variances (10^{-4}) of Js values

Method	Acc	Var (Acc)	Var (CSF)	Var (GM)	Var (WM)
FCN [2]	74.26	9.44	86.01	26.51	268.11
U-Net [1]	90.01	6.21	19.65	16.76	137.23
DRINet [3]	88.32	7.60	24.06	22.17	180.87
EDDNet [4]	88.82	5.84	18.98	17.31	167.56
MDUNet [12]	88.45	7.58	19.30	15.59	223.35
ResUNet-a [10]	90.43	4.80	19.10	14.83	117.14
U-Net++ [13]	87.22	12.02	32.52	23.10	203.71
TRMFCN (our)	91.81	3.86	14.76	13.63	126.07

4.2 Segmentation in 2.5D infant brain MR images

Data preprocessing: In order to acquire complete neighborhood information, we reserve all the black background images and obtain 2560 groups of images. These extra black background images are only used as domain information. The test set (96 groups) is same as the **4.1**.

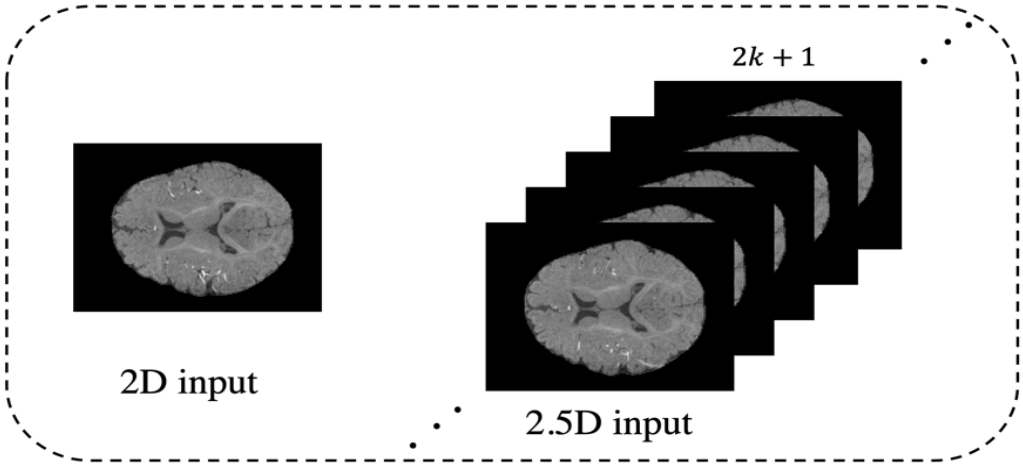

Fig. 7. The left half is the normal 2D input and the right half is the 2.5D input.

Model train: We changed the input of the model to a 2.5D form [28], which makes the network learn the domain feature information of the image to improve the segmentation accuracy. **Fig. 7** shows the difference between 2D input and 2.5D input. The 2D input has only one channel, but the 2.5D input has $2k + 1$ channels (From the above k images and the following k images), where k represents the number of the neighborhood images. In the experiment, we chose the value of k to be 2. Therefore, in our proposed method: the first input is five T1-w images ($192 \times 144 \times 5$), the second input is five T2-w images ($192 \times 144 \times 5$), and the third input is the concatenation of the first two inputs ($192 \times 144 \times 10$). The input form of all comparison methods is the same as the third input of our method. In network training, we still use the Adam optimizer to set mini-batch size to 10 and the iterations to 80. The dropout rate is set to 0.3 after the residual multiscale block.

Experiment Results: With the same number of iterations, the Js value of our method is the highest among all the eight methods, and **Table 3** shows the performance of each method visually. In **Table 1** and **Table 3**, compared to the 2D input form, the 2.5D input form in most models has better segmentation accuracy. In **Table 4**, although the variances of white matter and accuracy are not the lowest, in general, our model is still the most robust.

Table 3. Comparison of mean Js values of CSF, GM, WM% in eight 2.5D methods

Method(2.5D)	CSF (%)	GM (%)	WM (%)
FCN [2]	53.29	63.02	52.81
U-Net [1]	90.64	85.28	78.06
DRINet [3]	83.16	81.87	77.08
EDDNet [4]	88.00	82.84	75.54
MDUNet [12]	89.19	81.14	70.52
ResUNet-a [10]	88.97	84.49	78.09
U-Net++ [13]	83.95	73.63	60.07
TRMFCN (our)	91.10	85.88	79.55

Table 4. 2.5D Methods of accuracy (%) and variances (10^{-4}) of Js values

Method(2.5D)	Acc	Var (Acc)	Var (CSF)	Var (GM)	Var (WM)
FCN [2]	75.06	8.25	85.39	21.34	262.39
U-Net [1]	92.49	4.12	13.10	14.99	152.13
DRINet [3]	89.97	5.42	21.95	19.46	95.94
EDDNet [4]	90.99	4.89	13.01	17.72	163.95
MDUNet [12]	90.16	5.22	12.79	14.62	198.48
ResUNet-a [10]	92.06	3.60	12.93	14.28	130.15
U-Net++ [13]	85.42	13.67	21.81	33.40	268.78
TRMFCN (our)	92.93	3.72	12.77	14.06	122.27

5. Discussion and Conclusion

In this paper, we propose a full convolution network TRMFCN based on multi-modal data characteristics. We create a new form of input, residual multiscale (RM) block and concatenate block. The residual multiscale block in the structure solves the problem of gradient diffusion and makes the network training more efficient. The concatenate block greatly enhances the reusability of features to make the global feature information more fully learned. Our model is flexible for NMR multi-modal image data. If a new modal data is added, we can extend an extra branch.

Our method can also be applied to single-mode MR image segmentation. We selected a dataset of adult brain MR images from the Internet Brain Segmentation Repository (IBSR) to validate our ideas. There are 2304 adult brain MR images and 2304 corresponding labels in this data set. We also remove 257 images of all black backgrounds to optimize the data. Then we randomly selected 247 images as the test set and 1800 images as the training set, finally we used 200 images in the training set for verification. We changed all three input values to the same, so that a brain MR image can enter the network. We listed all Js values in **Table 5** after the test, where the results of TRMFCN are still the best.

Table 5. Comparison of mean Js values of CSF, GM, WM% in eight methods

Method(2.5D)	CSF (%)	GM (%)	WM (%)
FCN [2]	30.22	77.84	70.61
U-Net [1]	78.08	92.27	90.61
DRINet [3]	54.82	85.74	81.82
EDDNet [4]	68.26	89.97	87.35
MDUNet [12]	75.33	89.74	86.37
ResUNet-a [10]	74.70	91.27	89.35
U-Net++ [13]	43.83	71.73	66.98
TRMFCN (our)	78.41	92.82	91.25

Because of the extension of structure and the increase of the concatenated feature maps, the parameters will become more many. In order to reduce the amount of calculation, we will improve the method and even use the transfer learning [29] in the future. Meanwhile, we also guarantee its excellent ability.

References

[1] Ronneberger O, Fischer P, Brox T, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Proc. of MICCAI 2015. Springer International Publishing*, 234-241, 2015. Article (CrossRef Link).

[2] Long J, Shelhamer E, Darrell T, "Fully convolutional networks for semantic segmentation," *IEEE*, 3431-3440, 2015. Article (CrossRef Link).

[3] Chen L, Bentley P, Mori K, et al., "DRINet for Medical Image Segmentation," *IEEE Transactions on Medical Imaging*, 2018. Article (CrossRef Link).

[4] Chen L, Bentley P, Rueckert D, "Fully automatic acute ischemic lesion segmentation in DWI using convolutional neural networks," *Neuroimage Clinical*, 15, 633-643, 2017. Article (CrossRef Link).

[5] Huang G, Liu Z, Laurens V D M, et al., "Densely Connected Convolutional Networks," 2016. Article (CrossRef Link).

[6] He K, Zhang X, Ren S, et al., "Deep Residual Learning for Image Recognition," in *Proc. of IEEE Conference on Computer Vision & Pattern Recognition. IEEE Computer Society*, vol. 1, pp. 770-778, 2016. Article (CrossRef Link).

[7] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. of CVPR*, pp. 1–9, 2015. Article (CrossRef Link).

[8] Simonyan K, Zisserman A, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *Computer Science*, 2014. Article (CrossRef Link).

[9] Princy Matlani and Manish Shrivastava, "Hybrid Deep VGG-NET Convolutional Classifier for Video Smoke Detection," CMES: Computer Modeling in Engineering & Sciences, Vol.119, No.3, pp.427-458, 2019. Article (CrossRef Link).

[10] Diakogiannis F I, Waldner, François, Caccetta P, et al., “ResUNet-a: a deep learning framework for semantic segmentation of remotely sensed data,” 2019. Article (CrossRef Link).
[11] Lanlan Rui, Yabin Qin, Biyao Li and Zhipeng Gao, “Context-Based Intelligent Scheduling and Knowledge Push Algorithms for AR-Assist Communication Network Maintenance,” *CMES: Computer Modeling in Engineering & Sciences*, Vol.118, No.2, pp.291-315, 2019. Article (CrossRef Link).
[12] Zhang J, Jin Y, Xu J, et al., “MDU-Net: Multi-scale Densely Connected U-Net for biomedical image segmentation,” 2018. Article (CrossRef Link).
[13] Zhou Z, Siddiquee M M R, Tajbakhsh N, et al., “UNet++: A Nested U-Net Architecture for Medical Image Segmentation,” 2018. Article (CrossRef Link).
[14] Badrinarayanan V, Kendall A, Cipolla R, “SegNet: A Deep Convolutional Encoder-Decoder Architecture for Scene Segmentation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1-1, 2017. Article (CrossRef Link).
[15] Zhao H, Shi J, Qi X, et al., “Pyramid Scene Parsing Network,” 2016. Article (CrossRef Link).
[16] Chen L C, Papandreou G, Kokkinos I, et al., “Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs,” *Computer Science*, 2014(4), 357-361, 2014. Article (CrossRef Link).
[17] Lin G, Milan A, Shen C, et al., “RefineNet: Multi-Path Refinement Networks for High-Resolution Semantic Segmentation,” 2016. Article (CrossRef Link).
[18] G. Li, et al., "Early Diagnosis of Autism Disease by Multi-Channel Cnns," *Machine Learning in Medical Imaging*, pp. 303-309, 2018. Article (CrossRef Link).
[19] S. Hu, et al., "Learning-Based Deformable Image Registration for Infant MR Images in the First Year of Life," *Medical Physics*, vol. 44, pp. 158-170, Jan 2017. Article (CrossRef Link).
[20] F. Shi, et al., "Neonatal Atlas Construction Using Sparse Representation," *Human Brain Mapping*, vol. 35, pp. 4663-4677, Sep 2014. Article (CrossRef Link).
[21] F. Shi, et al., "Construction of Multi-Region-Multi-Reference Atlases for Neonatal Brain MRI Segmentation," *NeuroImage*, vol. 51, pp. 684-693, Jun 2010. Article (CrossRef Link).
[22] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” *ICML*, pp. 448–456, 2015. Article (CrossRef Link).
[23] Kazuhiko Kakuda, Tomoyuki Enomoto and Shinichiro Miura, “Nonlinear Activation Functions in CNN Based on Fluid Dynamics and Its Applications,” *CMES: Computer Modeling in Engineering & Sciences*, Vol.118, No.1, pp.1-14, 2019. Article (CrossRef Link).
[24] He K, Zhang X, Ren S, et al., “Identity Mappings in Deep Residual Networks,” 2016. Article (CrossRef Link).
[25] Li C, Xu C, Anderson A W, et al., “MRI Tissue Classification and Bias Field Estimation Based on Coherent Local Intensity Clustering: A Unified Energy Minimization Framework,” *Springer Berlin Heidelberg*, 288-299, 2009. Article (CrossRef Link).
[26] L. Wang, et al., "Longitudinally Guided Level Sets for Consistent Tissue Segmentation of Neonates," *Human Brain Mapping*, vol. 34, pp. 956-972, Apr 2013. Article (CrossRef Link).
[27] N. Srivastava, et al., "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *The Journal of Machine Learning Research*, vol. 15, pp. 1929-1958, 2014. Article (CrossRef Link).
[28] Hu K, Liu C, Yu X, et al., “A 2.5D Cancer Segmentation for MRI Images Based on U-Net,” in *Proc. of 2018 5th International Conference on Information Science and Control Engineering (ICISCE)*, 2018. Article (CrossRef Link).
[29] Zamir A, Sax A, Shen W, et al., “Taskonomy: Disentangling Task Transfer Learning,” 2018. Article (CrossRef Link).

Yunjie Chen received the B.S. and M.S. degree from the school of Math and Statistics, Nanjing University of Information Science and Technology, Nanjing, China, in 2002 and 2005, respectively. And the Ph.D. degree from the School of Computer Science and Engineering, Nanjing University of Science and Technology, in 2009. He is currently a Professor with the School of Math and Statistics, Nanjing University of Information Science and Technology. His research interests include medical image processing, statistical analysis, and machine learning.
E-mail: priestcyj@nuist.edu.cn

Yuhang Qin received the B.S. degree in Heilongjiang University of Science and Technology, Harbin, China, in 2019. His research interest is mainly focused on image segmentation with deep learning.
E-mail: qinyuhang98@163.com

Zilong Jin received the B.E. degree in computer engineering from Harbin University of Science and Technology, China, in 2009, and the M.S. and Ph.D. degrees in computer engineering from Kyung Hee University, Korea, in 2011 and 2016, respectively. He is currently an assistant professor of School of Computer and Software at Nanjing University of Information Science and Technology, China. His research interests include wireless sensor networks, mobile wireless networks, and cognitive radio networks.
E-mail: zljin@nuist.edu.cn

Zhiyong Fan received MSc. from Nanjing University of Information Science and echnology (NUIST) in 2007, China. He received Ph.D. degree in Nanjing University of Science and Technology in 2016, China. Now, he is a lecturer in the School of Automation at NUIST, China. His current research interests include medical imaging, image processing and pattern recognition.
E-mail: zhiyongfan1981@163.com

Mao Cai received the B.S. degree in Information and Computing Science from Nanjing University of Information Science and Technology, Nanjing, China, in 2019. His research interest is mainly focused on pattern recognition, image segmentation, and image processing.
E-mail: 1440896206@qq.com