F2RNET: A FULL-RESOLUTION REPRESENTATION NETWORK FOR BIOMEDICAL IMAGE SEGMENTATION

*Junlong Cheng, Chengrui Gao, Changlin Li, Zhangqiang Ming, Yong Yang, Fengjie Wang, Min Zhu**

College of Computer Science, Sichuan University, Chengdu 610065, China

2022 IEEE International Conference on Image Processing (ICIP) | 978-1-6654-9620-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICIP46576.2022.9897649

ABSTRACT

In this paper, we are interested in exploring the problem of full-resolution image segmentation, with the focus placed on learning full-resolution representations for biomedicine images. We divide the original resolution image into patches of different sizes in different stages and then extracte local features from large to small patches using efficient and flexible components in modern convolutional neural networks (CNN). Meanwhile, a multilayer perceptron (MLP) block intended for modeling long-range dependencies between patches is designed to compensate for the inherent inductive bias caused by convolution operations. In addition, we perform multi-scale fusion and receive representation information from parallel paths at each stage, resulting in a rich full-resolution representation. We evaluate the proposed method on different biomedical image segmentation tasks and it achieves a competitive performance compared to the latest deep learning segmentation methods. It is hoped that this method will serve as a useful alternative to biomedical image segmentation and provide an improved idea for the research based on full-resolution representation.

***Index Terms*—** Full-resolution representation, Local feature, Multi-scale fusion, Biomedical image segmentation

1. INTRODUCTION

Biomedical image segmentation plays a key role in computer-aided diagnosis, which is aimed to extract the regions of interest in an image. With the development of medical imaging technology, the sample size and diversity of biomedical images are rapidly increasing so that manual segmentation is no longer sufficient to meet practical needs. Therefore, it is important to develop automatic, accurate and robust biomedical image segmentation methods. In the case of semantic segmentation, state-of-the-art approaches rely on encoder-decoder architectures. For example, U-Net [1] and others use encoder networks to learn high-level semantic representations and decoders for the recovery of lost spatial information from the high-level representations. DeepLabv3 [2] expands the receptive field and aggregates multi-scale information by atrous convolution and pooling operations. Some works [3, 4] introduce built-in depth variable U-Net ensembles and redesign skip connections to achieve more flexible feature fusion. Besides, enhancing the capability of feature representation for the encoder and decoder [5] is a means to improve the performance in segmentation. Another approach is to utilize a self-attention mechanism to generate more discriminative feature representations and model the long-range dependencies of input features [6, 7]. TransUNet [8] and U-Net Transformer [9] learn patches sequence relations through Transformer. Transformer-based methods usually work well when trained on large-scale datasets [10], but the number of images available for training in medical datasets tends to be relatively small. In summary, reducing the loss of feature information, fusing multi-scale features and performing well on small datasets are the key issues that need to be addressed for biomedical image segmentation.

We propose a new full-resolution representation network to solve the above-mentioned problems. Compared with those networks [1, 3, 8, 11] widely used now for biomedical image segmentation, our network has three benefits: (i) We maintain a full resolution representation throughout, thus avoiding the problem of losing image detail information through down-sampling. (ii) Unlike the ordinary segmentation networks ensuring that richer feature information is extracted by increasing the number of channels at each stage, we maintain an competitive segmentation performance while using the same number of channels at each stage. (iii) Most existing methods have the feature fusion schemes that fuse shallow and deep features (e.g., U-Net and its variants). Differently, we perform multi-scale fusion at the same depth (stage) to improve the full-resolution representation at the different stages.

2. RELATED WORK

In theory, a well-performing segmentation network should try to keep as much detail as possible in the image while reducing redundant features. Learning a full-resolution or high-resolution representation of an image can be a solution. For example, Li et al. [12] used atrous convolution and residual concatenation to achieve an end-to-end mapping from image volume to voxel-level dense segmentation. Mohammed et al. [13] increased the depth of the network while maintaining the full-resolution to simulate changes in the receptive field of down-sampling. The design of multi-branch network can

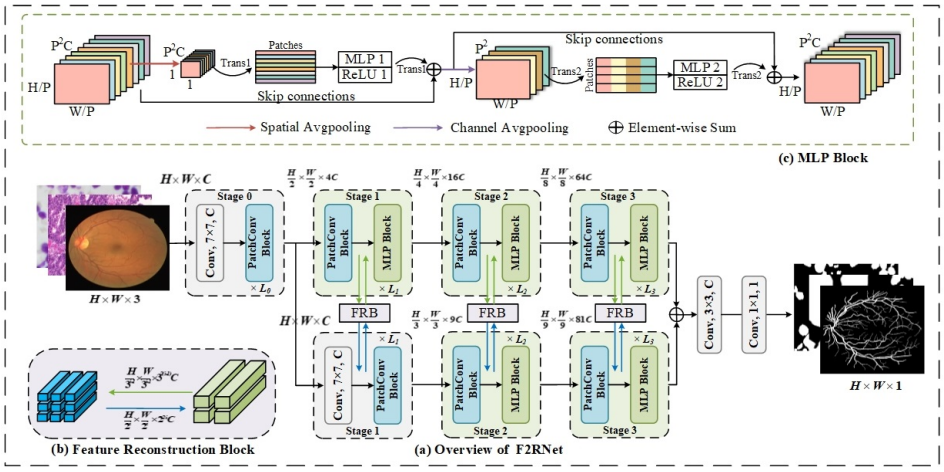

Fig. 1. The framework of our proposed method: (a) overview of F2RNet; (b) Feature reconstruction block; (c) MLP block

meet the requirement of feature complementarity in downstream tasks, and Qu et al. [11] learned basic semantic information and non-destructive semantic information through U-shaped branch and full feature extraction branch.

A simple multi-scale fusion approach is to separately feed the input features into parallel filters of different kernel sizes and aggregate the feature maps of different outputs. This multi-branch approach is widely used by GoogleNet [14]. U-Net and its variants [3, 15, 16] gradually fuse shallow features from high to low resolution into deeper features at the same resolution from low to high resolution.

Attentional modules can model long-range dependencies and have been widely used in various visual tasks [7, 17, 18]. Before vision transformers (ViT) were proposed, attention was either used with CNN in conjunction or used to replace some components of CNN. ViT can also perform image classification tasks well with a pure transformer training sequence of image patches directly. Chen et al. [8] encode tokenized image patches in CNN feature maps as input sequences, and design TransU-Net with the advantages of both transformer and U-Net. Extensions of the above work also include [9, 19].

3. METHODOLOGY

An overview of our proposed F2RNet for biomedical image segmentation is shown in Figure 1 (a). The framework is mainly composed of three parts, including PatchConv Block, Feature Reconstruction Block and MLP Block.

3.1. PatchConv Block

The essence of the PatchConv block is to perform a convolution operation on different patches to extract local features of the image. We denote the input feature map of this module by $M_{in} \in R^{N\times C\times H\times W}$, where N is the batch size, C is the number of channels, H and W are the height and width, respectively. First split M_{in} into a series of patches of size $(\frac{H}{2^i}, \frac{W}{2^i})$ and $(\frac{H}{3^{i-1}}, \frac{W}{3^{i-1}})$, "i" represents the number of stages of the module $(i \geq 0)$. The number of these patches is gradually increased, and the resolution is 1/2 or 1/3 of the previous stage (depending on which branch). The operation of split into patches does not move data in memory and does not perform training. It can be done by the following three steps (the 1st branch is used as an example below): 1) Reshape M_{in} to $(N, C, 2^i, \frac{H}{2^i}, 2^i, \frac{W}{2^i})$; 2) Rearrange the order of the axes as $(N, 2^i, 2^i, C, \frac{H}{2^i}, \frac{W}{2^i})$; 3) Reshape the above features to $M_p \in R^{N\times 2^{2i}\times C\times \frac{H}{2^i}\times \frac{W}{2^i}}$ again. Then, act on M_p with a set of shared convolutions with residual connections, which can be expressed as:

$$M_{conv} = \left[M_p^n + Conv(M_p^n, C, K, D, P)\right] \qquad (1)$$

where $M_{conv} \in R^{N\times 2^{2i}C\times \frac{H}{2^i}\times \frac{W}{2^i}}$, $n \in [0, 1, ..., 2^{2i}]$, [...] indicates connected along the channel direction, C is the number of output channels of convolution, K is the convolution kernel size (default is 3), D is the dilation rate, which doubles as the number of cycles increases, and P is the number of pixels to be filled. Finally, use layer normalization to normalize M_{conv} to get $M_{out} \in R^{N\times 2^{2i}C\times \frac{H}{2^i}\times \frac{W}{2^i}}$, The reason we do not use traditional batch normalization here is that it will destroy the overall information of the image, resulting in reduced segmentation accuracy.

3.2. Feature Reconstrucation and MLP Block

The Feature reconstruction block (FRB) is designed to fuse the features of two branches and integrate semantic information at different scales. Figure 1(b) is a schematic diagram of the mutual conversion of two branches. When in the same stage, the tensors of the two branches are $M_{one}^i \in R^{N\times 2^{2i}C\times \frac{H}{2^i}\times \frac{W}{2^i}}$ and $M_{two}^i \in R^{N\times 3^{2(i-1)}C\times \frac{H}{3^{i-1}}\times \frac{W}{3^{i-1}}}$ respectively. For ease of fusion, we first convert the shapes of

M^i_{one} and M^i_{two} to $M^i_{one \longrightarrow two}$ and $M^i_{two \longrightarrow one}$ in the opposite order of the previous steps. Then, the four tensors are added element-wise, namely $M^i_{out1} = M^i_{one} + M^i_{two \longrightarrow one}$ and $M^i_{out2} = M^i_{two} + M^i_{one \longrightarrow two}$. As the stage gets deeper, the resolution of each image patch is gradually smaller and the local features are further refined, which are in line with human visual perception of images. Compared with shallow and deep feature fusion methods, our method can obtain more complementary information.

MLP block is shown in Figure 1(c), the resolution of original input images is (H, W) and the resolution of each patch is $(\frac{H}{P}, \frac{W}{P})$ (P denotes the number of patches in the length or width direction), the number of these patches is P^2 and each patch projection dimension is C.

The MLP block consists mainly of two multi-layer perceptron layers and non-linear layers. The first MLP layer acts on the spatially compressed patches and is used to learn the connection information between the projection channels of the different patches. Specifically, we first aggregate the spatial information of the input tensor $M_{in} \in R^{N \times \frac{H}{P} \times \frac{W}{P} \times P^2 C}$ using a spatial average pooling (SAP), and the pooled tensor is $M_{sap} \in R^{N \times 1 \times 1 \times P^2 C}$. After that, the M_{sap} is fed into the linear layer with the following transformation ("Trans 1"):

$$M_{Trans1} = Permute\left(Up(M_{sap})\right) \tag{2}$$

Which Up means that the patches in M_{sap} are up-sampled according to the pixel distribution of the original image, i.e., $R^{N \times 1 \times 1 \times P^2 C} \longrightarrow R^{N \times P \times P \times C}$ and then the tensor is reshaped to obtain M_{Trans1} of shape (NP^2, C).

The above transformation remains cost-free, and the input and output of MLP layer remain the same. Compared with direct linear mapping on the original input images or features, the required computation of our method is reduced from HWC^2 to P^2C^2. Lastly, we utilize the inverse operation of "Trans 1" to restore features after the nonlinear layer (ReLU) and fuse restored features with the input features.

The second MLP layer acts on the patches after channel compression and is used to learn the long-range relationships between patches. As described above, the channel information of the input tensor is first aggregated using channel average pooling (CAP), i.e., $M_{cap} \in R^{N \times \frac{H}{P} \times \frac{W}{P} \times P^2}$. And then M_{cap} is reshaped into a tensor of $(P^2, \frac{HW}{P^2})$ (i.e., "Trans 2" operation), each line of which contains all the information of one patch.

The second MLP layer receives inputs of different sizes at different stages, but the computational parameters remain the same (ignoring bias). As the stage gets deeper, this layer receives more image patches, this design is similar to the traditional pyramid structure, but we always learn the full-resolution representation of the image. Finally, we use the skip connection to conduct the Hadamard product for the results of the ReLU layer and the input features to obtain the output features of the MLP block.

3.3. F2RNet Architecture

The main body of our F2RNet contains four stages with two parallel sub-networks, where the 1-st stage extracts features from the input image using a convolution with the kernel of 7×7 and a PatchConv block. Starting from the 2-nd stage, we apply PatchConv block, Feature reconstruction block (FRB) and MLP block for multi-scale feature fusion of parallel paths and the generation of rich full-resolution representations.

We have built the base model F2RNet-B, and the parameter settings of this model are similar to ResNet-based backbone network. Furthermore, F2RNet-T, F2RNet-S and F2RNet-L are presented, as the complexity of these instances increases, so does performance. The architectural parameters of these model variants include:

$F2RNet - T : C = 32, L = [1, 1, 1, 1]$
$F2RNet - S : C = 32, L = [1, 2, 2, 2]$
$F2RNet - B : C = 64, L = [1, 1, 2, 2]$
$F2RNet - L : C = 96, L = [1, 1, 1, 2]$

Where C is the number of feature channels in the first stage, which remains constant throughout the network. L represents the number of cycles in different stage.

4. EXPERIMENTS

We conduct experiments on three public biomedical image segmentation datasets to verify the effectiveness of our method. The three datasets are the Kaggle 2018 data science bowl (referenced to as Nuclei) [7], Retinal Images vessel Tree Extraction (RITE) [20] and GLAnd Segmentation (GLAS) [20] datasets, and their image counts are 670, 160 and 165 respectively (The training resolution of F2RNet is 216 x 216, others are 224 × 224). To make our experimental results more convincing, we conduct all experiments using 5-fold cross-validation.

4.1. Implementation Details

We implement the Pytorch based method by training on NVIDIA Tesla V100 GPU. During training, we uses the Adam optimizer, the learning rate is fixed at 1e-4, the batch size is set to 16, and the Crossentropy loss function is adopted at the end of the network. When the validation loss is stable and there is no significant change within 30 epochs, an early stopping mechanism is used to stop training. The training data is augmented by applying random rotations (±25°), random horizontal, vertical shifts (15%), and random flips (horizontal and vertical). All comparative experiments are performed under the same operating environment, hyper-parameters, training set and validation set.

4.2. Comparison with State-of-the-arts

We compare the proposed F2RNet with eight state-of-the-art methods (U-Net [1], U-Net++ [3], ResUNet [5], R2UNet

Table 1. Comparison results with state-of-the-art methods on three datasets.

Method	Nuclei		GLAS		RITE	
	IoU	Dice	IoU	Dice	IoU	Dice
U-Net	84.20	91.42	85.28	92.04	68.08	81.02
U-Net++	84.54	91.64	86.12	92.56	66.80	80.08
ResUNet	85.48	92.18	85.88	92.38	69.82	82.20
R2UNet	85.76	92.18	85.90	92.52	68.38	81.06
BiONet	85.22	92.06	85.44	91.90	65.06	78.82
ResGANet	85.50	91.82	85.10	91.66	66.20	79.38
TransUNet	85.18	92.02	86.52	92.86	68.58	81.34
SwinUNet	84.76	91.76	83.08	90.74	67.80	80.82
F2RNet-T	85.24	91.60	84.86	91.50	69.84	82.04
F2RNet-S	85.66	91.88	85.88	92.14	70.30	82.38
F2RNet-B	86.08	92.14	86.56	92.48	**70.42**	**82.46**
F2RNet-L	**86.29**	**92.26**	**87.12**	**92.88**	70.30	82.38

[16], BiONet [15], ResGANet [21], TransUNet [8] and SwinUNet [22]) on three datasets. We adopt the two most commonly used evaluation metrics in semantic segmentation (i.e., IoU and Dice) to evaluate above methods.

We can observe from Table 1. that the performance of F2RNet improves as the model size gets larger generally. It is worth noting that our method is trained from scratch without relying on any pre-trained models, which indicates that our method is less dependent on the amount of data. On the RITE dataset, F2RNet-B obtained the best performance. The reason for the inferior performance of F2RNet-L on this data may be due to the small size of the dataset and the sparse target. Smaller models are already able to learn important features needed to segment objects. The IoU of our method is 3.64%-4.22% higher than that of the attention-based method ResGANet. Moreover, these experimental results illustrate the advantages of our framework in extracting sparse features compared to previous methods. The IoU and Dice of F2RNet-L on the GLAS dataset are 87.12% and 92.88%, respectively, which are better than current Transformer-based methods. These results demonstrate the good generalization ability of our method, and also illustrate the potential of full-resolution representations in the field of semantic segmentation.

4.3. Analytical Study

To comprehensively evaluate the proposed F2RNet framework, we conduct multiple ablation studies, including ablation studies of MLP blocks, ablation studies of different branches and feature fusion.

On the influence of MLP block. As mentioned previously, MLP blocks can improve the detail of semantic segmentation by modeling long-range dependencies between patches. Table 2. shows the segmentation results on the Nuclei and RITE datasets, and we can see that the addition of the MLP block to F2RNet generally leads to better segmentation performance. Even on smaller datasets like RITE, there is still a consistent improvement. These experimental results reinforce our initial intuition for designing the MLP, and we believe that training the MLP block on a larger dataset is likely to have better performance.

Table 2. Ablation studies of MLP block on Nuclei and RITE datasets.

Method	Nuclei		RITE	
	IoU	Dice	IoU	Dice
F2RNet-T	86.30 (↑0.74)	92.71 (↑0.15)	68.31 (↑0.61)	80.96 (↑0.37)
F2RNet-S	86.61 (↑0.63)	92.80 (↑0.15)	68.30 (↑1.16)	80.91 (↑0.84)
F2RNet-B	87.02 (↑0.65)	93.03 (↑0.26)	69.72 (↑0.71)	81.90 (↑0.70)
F2RNet-L	87.30 (↑0.76)	93.11 (↑0.38)	69.44 (↑1.15)	81.77 (↑0.95)

Table 3. Results of the ablation study on the GLAS dataset.

Branch 1	Branch 2	Fusion	IoU (%)	Dice (%)
√			87.21	93.10
	√		86.83	92.84
√	√		88.29	93.62
√	√	√	**88.97**	**94.19**

On the influence of different branches and feature fusion. Our method performs multiscale fusion between two different branches at the same depth aiming to improve the full-resolution representation of the different stages. We used F2RNet-B as a benchmark on the GLAS dataset to verify whether different branches and feature fusions affect segmentation performance. It can be observed from Table 3. that "Branch 1", which divides image patches with even numbers, has better overall performance than "Branch 2", which divides patches with odd numbers, but "Branch 2" obtains the highest precision index. The 3-rd row shows the experimental results for two branches without feature interaction and fusion. Using two branches to extract features can improve IoU by 1.0% to 1.4%. The best experimental results are obtained when the two branches and feature fusion operations are used together. This also demonstrates that the improvements we have made in the model are beneficial in improving medical image segmentation performance.

5. CONCLUSION

We explore full-resolution representations for biomedical image segmentation in this work. First, grouped convolutions in spatial dimensions are utilized to act on pyramid-structured patches for learning local features of images. Then, we propose an MLP block based on a multi-layer perceptual architecture to enhance the long-range dependencies between different patches. Finally, rich full-resolution representations are generated by multi-scale feature fusion at the same depth. The experimental results demonstrate the superiority of our proposed method over the current ConvNet and Transformer-based approaches.

6. REFERENCES

[1] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*. 2015, vol. 9351, pp. 234–241, Springer.

[2] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam, "Rethinking atrous convolution for semantic image segmentation," *CoRR*, vol. abs/1706.05587, 2017.

[3] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *DLMIA*. 2018, vol. 11045, pp. 3–11, Springer.

[4] Xuebin Qin, Zichen Vincent Zhang, Chenyang Huang, and Masood Dehghan, "U^2-net: Going deeper with nested u-structure for salient object detection," *Pattern Recognit.*, vol. 106, pp. 107404, 2020.

[5] Xiao Xiao, Shen Lian, Zhiming Luo, and Shaozi Li, "Weighted res-unet for high-quality retina vessel segmentation," 10 2018, pp. 327–331.

[6] Xinze Li, Bangyu Wu, Xu Zhu, and Hui Yang, "Consecutively missing seismic data interpolation based on coordinate attention unet," *IEEE Geosci. Remote. Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[7] Junlong Cheng, Shengwei Tian, and Long Yu, "Fully convolutional attention network for biomedical image segmentation," *Artif. Intell. Medicine*, vol. 107, pp. 101899, 2020.

[8] Jieneng Chen, Yongyi Lu, Qihang Yu, and Xiangde Luo, "Transunet: Transformers make strong encoders for medical image segmentation," *CoRR*, vol. abs/2102.04306, 2021.

[9] Olivier Petit, Nicolas Thome, and Clement Rambour, "U-net transformer: Self and cross attention for medical image segmentation," in *MICCAI*. 2021, vol. 12966, pp. 267–276, Springer.

[10] Alexey Dosovitskiy, Lucas Beyer, and Alexander Kolesnikov, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR, 2021*, pp. 3–7.

[11] Lei Qu, Meng Wang, Kaixuan Guo, and Wan Wan, "Biomedical image segmentation based on full-resolution network," *Pattern Recognit. Lett.*, vol. 153, pp. 232–238, 2022.

[12] Wenqi Li, Guotai Wang, and Lucas Fidon, "On the compactness, efficiency, and representation of 3d convolutional networks: Brain parcellation as a pretext task," in *IPMI*. 2017, vol. 10265, pp. 348–360, Springer.

[13] Mohammed A. Al-masni, Mugahed A. Al-antari, and Mun-Taek Choi, "Skin lesion segmentation in dermoscopy images via deep full resolution convolutional networks," *Comput. Methods Programs Biomed.*, vol. 162, pp. 221–231, 2018.

[14] Christian Szegedy, Wei Liu, and Yangqing Jia, "Going deeper with convolutions," in *CVPR*. 2015, pp. 1–9, IEEE Computer Society.

[15] Tiange Xiang, Chaoyi Zhang, and Dongnan Liu, "Bio-net: Learning recurrent bi-directional connections for encoder-decoder architecture," in *MICCAI*. 2020, vol. 12261, pp. 74–84, Springer.

[16] Md. Zahangir Alom, Mahmudul Hasan, and Chris Yakopcic, "Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation," *CoRR*, vol. abs/1802.06955, 2018.

[17] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, 2020.

[18] Xin Ning, Ke Gong, and Weijun Li, "Feature refinement and filter network for person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 9, pp. 3391–3402, 2021.

[19] Huisi Wu, Shihuai Chen, and Guilian Chen, "Fat-net: Feature adaptive transformers for automated skin lesion segmentation," *Medical Image Anal.*, vol. 76, pp. 102327, 2022.

[20] Jeya Maria Jose Valanarasu, Vishwanath A. Sindagi, and Ilker Hacihaliloglu, "Kiu-net: Towards accurate segmentation of biomedical images using over-complete representations," in *MICCAI*. 2020, vol. 12264, pp. 363–373, Springer.

[21] Junlong Cheng, Shengwei Tian, and Long Yu, "Resganet: Residual group attention network for medical image classification and segmentation," *Medical Image Anal.*, vol. 76, pp. 102313, 2022.

[22] Hu Cao, Yueyue Wang, and Joy Chen, "Swin-unet: Unet-like pure transformer for medical image segmentation," *CoRR*, vol. abs/2105.05537, 2021.