

FEATURE FUSION ENCODER DECODER NETWORK FOR AUTOMATIC LIVER LESION SEGMENTATION

Xueying Chen¹, Rong Zhang¹, Pingkun Yan²

¹Department of Electronic Engineering and Information Science,
University of Science and Technology of China, Hefei, China

²Department of Biomedical Engineering,
Rensselaer Polytechnic Institute, Troy, NY 12180, USA

ABSTRACT

Liver lesion segmentation is a difficult yet critical task for medical image analysis. Recently, deep learning based image segmentation methods have achieved promising performance, which can be divided into three categories: 2D, 2.5D and 3D, based on the dimensionality of the models. However, 2.5D and 3D methods can have very high complexity and 2D methods may not perform satisfactorily. To obtain competitive performance with low complexity, in this paper, we propose a Feature-fusion Encoder-Decoder Network (FED-Net) based 2D segmentation model to tackle the challenging problem of liver lesion segmentation from CT images. Our feature fusion method is based on the attention mechanism, which fuses high-level features carrying semantic information with low-level features having image details. Additionally, to compensate for the information loss during the upsampling process, a dense upsampling convolution and a residual convolutional structure are proposed. We tested our method on the dataset of MICCAI 2017 Liver Tumor Segmentation (LiTS) Challenge and achieved competitive results compared with other state-of-the-art methods.

Index Terms— Liver lesion segmentation, deep learning, encoder-decoder, attention, feature fusion.

1. INTRODUCTION

Liver cancer is one of the most commonly seen cancers with high mortality rate in the modern world. Computed tomography (CT) is often used for tumor diagnosis and treatment. Manual liver lesion segmentation on CT is time-consuming, difficult to reproduce, and easily influenced by personal subjective experience. Therefore, fully automatic method for liver lesion segmentation needs to be developed. However, automatic liver lesion segmentation is a very challenging task due to the following reasons. First, CT images often have low contrast for soft tissues and the imaging noise makes it difficult to segment the lesions. Second, the position and the shape of the lesion vary significantly. Last but not the least, the lesion boundaries are unclear and the sizes of most lesions are relatively small.

To solve the problems mentioned above, a number of methods based on deep learning have been proposed for the liver lesion segmentation task in the past several years. Generally, these learning methods can be divided into three categories: 1) 2D models, such as Cascaded-FCN [1], FCN based on VGG-16 [2]; 2) 2.5D models, such as U-Net using residual connections [3] and method exploiting inter-slice similarity [4]; 3) 3D models, such as Densely connected volumetric ConvNets [5], H-DenseUNet [6], 3D FCN [7]. Even though the previous methods have obtained much improved perfor-

mance, there still exist a lot of potentials. For example, previous 2D networks cannot fully utilize the depth information of the CT image and thus its segmentation accuracy is low. On the other hand, 2.5D and 3D networks are computationally expensive and have very high requirement for hardware configurations. Given the same computing resources, the large amount of parameters and heavy computation limit the depth and complexity of the 3D networks.

With the aforementioned considerations, we propose a Feature-fusion Encoder-Decoder Network (FED-Net) based 2D model for image segmentation. First, we design a novel feature fusion method based on the attention mechanism, which can effectively embed more semantic information into low-level features and improve the current feature fusion mode based on the U-Net architecture network. Then, to compensate for the information loss during the upsampling process, we propose a dense upsampling convolution instead of using the traditional upsampling operations and also add residual convolutional blocks to refine the rough boundary of the target.

2. METHOD

Among the many aforementioned problems of convolutional neural network (CNN) based image segmentation methods, one major issue with semantic segmentation is due to the pooling layers. Pooling layer not only expands the receptive field but also aggregates the context while discarding the location information. However, semantic segmentation requires exact alignment of class maps and thus, needs the location information to be preserved. To solve such problem, many encoder-decoder architectures have been proposed. In the encoder-decoder framework, the encoder gradually reduces the spatial dimension of the feature maps and the decoder gradually recovers the object details and spatial dimensions. Skip connections are usually used between the encoder and decoder, which help the decoder to better retain the details of the target, especially for situations with small sample sizes. The widely adopted U-Net architecture [8] is a typical encoder-decoder structure. There are quite a few models that fuse features in this fashion, for example RefineNet [9] and ExFuse [10]. However, typical U-Net based networks directly fuse high-resolution features from encoder path to the upsampled decoder output through either addition, multiplication, or concatenation, which may not be able to efficiently use features from different levels. Our work suggests that all levels of features should contribute more directly to semantic segmentation and the importance of features should be weighed differently.

In this paper, we propose a new method for feature fusion with an attention mechanism. To refine the rough boundary due to up-

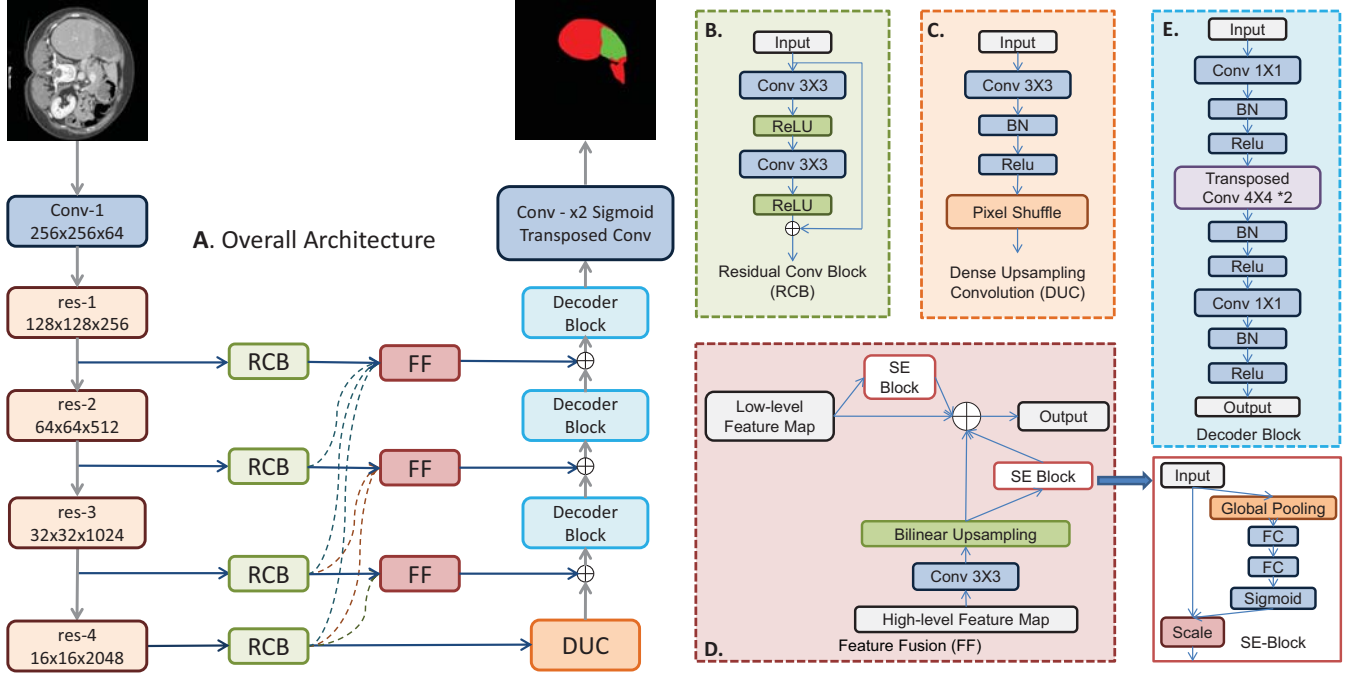


Fig. 1. The overall architecture of our proposed approach and the individual components of the FED-Net. The encoder part uses ResNet-50 with the fully connected layer removed. **RCB** - Residual Convolution Block, **FF** - Feature Fusion Based on Attention, **DUC** - Dense Upsampling Convolution. **Decoder block** - first reduce the number of filters by 1×1 convolution, then perform transposed convolution, and finally restore the number of filters by 1×1 convolution.

sampling operation, residual convolution blocks (RCBs) are inserted before feature fusion. Then after the feature fusion, in order to make up for the lost location information during the upsampling process, the dense upsampling convolution (DUC) is applied. An overview of the proposed method is given in the left side of Fig. 1.

2.1. FED-Net

In this subsection, details of the proposed Feature-fusion Encoder-Decoder Network (FED-net) for medical image segmentation are provided. The encoder part use pre-trained model on the ImageNet - ResNet50 [11]. We divide the pre-trained ResNet-50 into four blocks according to the resolution of the feature map, and add the residual convolution block after the four blocks separately. Feature fusion is then performed on the output of each block, and finally each unit is sent to the decoder part.

Attention based Feature Fusion. For convolutional neural networks, the low-level general extracts the low-level features and contains little semantic information, which has been confirmed by ZF-Net [12]. As mentioned above, the typical U-Net based networks directly fuse high-resolution features from encoder path to the up-sampled decoder output, which may not be able to efficiently use features from different levels. Inspired by SE-NET [13], we proposed a feature fusion method based on attention mechanism to fuse features of different levels. We generalize the fusion method into the form of

$$H_l = SE(x_l) + \sum_{i=l+1}^L SE((U(x_i))) \quad (1)$$

where L is the number of feature levels. The feature fusion method

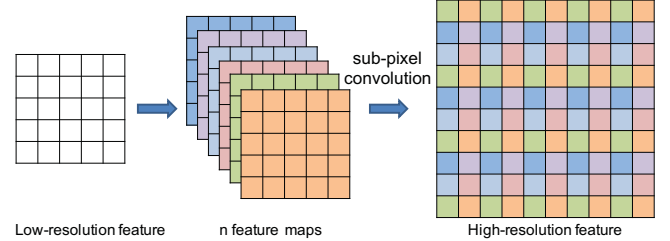


Fig. 2. Sub-pixel convolutional module with pixel shuffling.

is illustrated in Fig. 1-D. Our feature fusion first extracts multi-resolution features from all levels of the encoder. Then the feature fusion module at each level fuses the features at its current level and also the levels above it. Where the function $U(\cdot)$ - upsampling operation is performed on the high-level features, and the SE-Net is executed on all the features that need to be fused. The detail of SE-block are shown in the lower right corner of Fig. 1. With this feature fusion approach, more useful features can be extracted by assigning different weights to the features from different based on their contribution to the final segmentation results. Then, the importance of those features which have more impact on the segmentation results can be emphasized.

Residual Convolution Block. A residual convolution block (RCB) that does not include batch normalization is added after the four different resolution blocks that are extracted from the encoder part. The RCB is easy to train using the “shortcut connection” method and further extracts intermediate features to improve the results. The details of the RCB structure is shown in Fig. 1-B.

Dense Upsampling Convolution. To compensate for the details lost during the upsampling process, dense upsampling convolution (DUC) blocks are used which is a sub-pixel convolution with pixel shuffling [14]. The structure of DUC is in Fig. 1-C. The details of pixel shuffling are shown in Fig. 2, which is mainly to periodically merge multiple low-resolution feature maps into a high-resolution map according to specific positions determined by a “period shuffling” operation. We use DUC as the first block of the decoder part. That is to have the highest level feature map not only contain highly abstract features but also stay in relatively higher resolution. This is designed to deal with the commonly seen block artifact from fully convolutional neural networks based segmentation methods.

2.2. Loss Function

Considering the imbalance between the positive and negative samples of the data set, we combined the binary Cross Entropy with the Jaccard index. The overall loss of our proposed method is given below:

$$L(y, \hat{y}) = (\omega_1 - 1) * y \log \hat{y} - \omega_1 * (1 - y) \log(1 - \hat{y}) - \omega_2 * \log\left(\frac{|y \cap \hat{y}| + \epsilon}{|y| + |\hat{y}| - |y \cap \hat{y}| + \epsilon}\right) \quad (2)$$

where y represents the ground truth segmentation of a slice and \hat{y} denotes the predicted mask of the slice.

The first two terms on the right of Eqn. (2) define a weighted binary cross entropy. The weighting parameter ω_1 can help alleviate the problem of data imbalance between positive and negative samples. In our application, the volume of lesion is quite small compared to the rest of the imaging volume. The third term of Eqn. (2) contains the generalized Jaccard index that is commonly used to measure the quality of segmentation result by computing the similarity between two segmentation as the size of the intersection divided by the size of the union. The generalized Jaccard index is used for two reasons. On the one hand, the index is differentiable, where back propagation can be conveniently computed. On the other hand, it is a proper distance metric relative to the Dice index. In the third term on the right side of the above equation, ϵ is a very small number to prevent the denominator being zero. In our work, we used $\epsilon = 1 \times 10^{-15}$.

2.3. Hierarchical Segmentation

For liver tumor segmentation, we take a two-step process. The first step is to segment the liver with a baseline segmentation network and the second step is to segment the lesion with the FED-Net. Our baseline network is a “U-Net” network using the ResNet-50 [11] as the encoder part.

In the first stage of the liver segmentation, we used the entire volume for training and validation. In the second stage for lesion segmentation, in order to reduce the number of negative samples and the amount of computation, we sent a slice containing only the liver into the network. In the test phase, only the liver slice obtained from the first step segmentation were used. To alleviating the imbalance between the positive and negative samples of the dataset, we implemented a probability extraction strategy to extract positive and negative samples from the dataset during training. The positive samples were extracted with a probability of 0.9, and the negative samples were extracted with a probability of 0.1.

To make full use of the third dimension information of the CT image and without adding extra calculations, we sent the adjacent three slices as the three channels of the intermediate slice into the network.

3. EXPERIMENTS AND RESULTS

3.1. Dataset and Preprocessing

We used the publicly available dataset of MICCAI 2017 LiTS Challenge as the training of our proposed method. The LiTS datasets contains 131 and 70 contrast-enhanced 3D abdominal CT scans for training and testing, respectively. We tested our method on the 70 LiTS challenge test cases. The dataset was acquired from a number of different clinical sites with different scanners and protocols. Each slice in all CT scans has a fixed size of 512×512 pixels with the number of slices per scan varies from 42 to 1026. The image resolutions are also different from scan to scan.

In order to remove irrelevant information about other organs and tissues in the CT scans for liver lesion segmentation, we cut the the image intensity values of all CT scans to the range of $[-200, 250]$ HU. After all CT scan HU values were truncated, we normalized all slice intensities into the range $[0, 1]$ with min-max normalization.

3.2. Implementation Details

Our proposed FED-Net was implemented using the pytorch package. Due to its efficiency, the experiment used a single NVIDIA GTX 1080 GPU with 8 GB memory. We used stochastic gradient descent method for optimization with momentum of 0.9 and weight decay of 0.0001. We used the CT scans from 0 to 119 and 120 to 130 of the LiTS dataset as our training set and validation set, respectively. Simple data augmentation including horizontal flipping and vertical flipping were performed.

In the post-processing stage: (1) The threshold for the output of liver segmentation is 0.5, and 3D connect-component labeling was used. (2) The threshold for the output of lesion segmentation is 0.3. On the basis of (1), a bounding-box was taken for the result of liver segmentation, and then the liver segmentation result with bounding-box was merged with the lesion segmentation result to obtain the final lesion segmentation result.

3.3. Ablation Analysis of FED-Net

In this subsection, we will make step-by-step comparisons to evaluate our approaches proposed in Section 2. In Section 2, we propose Attention based Feature Fusion (FF) to effectively assign weights to embed more semantic information into low-level features. To compensate for the missing details during the upsampling process, we propose Dense Upsampling Convolution (DUC) and Residual Convolutional Block (RCB). To demonstrate the effectiveness of our method, we tested the performance of each component separately. The results are presented in Table 1. From the results, we can find that the performance has increased correspondingly with the addition of FF, DUC and RCB. When we used the feature fusion method without SE-Block, we found that the performance hardly improve, and when we used our feature fusion method with SE-Block, the result increase by 0.8%.

We compare our results to a number of excellent methods, Table 2 shows the details. Most of the first few methods are based on 2.5D or 3D models, among them, H-DenseUNet and 3D AH-Net are 3D models, Res-UNet and FCN+Detector are 2.5D models, and Cascaded ResNet is 2D models, our model shows very promising performance. It isn't too far away from other leading 3D and 2.5D methods, given it is in 2D and the number of parameters is much less. It is light weighted and can be easily deployed. We hope to further improve the performance in future by extending it to 3D and also with other enhancements.

Table 1. Liver lesion segmentation Dice scores of our method on the LiTS challenge test cases under different ablations.

Model	Per case	Global
Baseline	0.635	0.745
Baseline + RCB	0.644	0.776
Baseline + FF	0.636	0.756
Baseline + FF with SE-Block	0.643	0.775
Baseline + DUC	0.644	0.772
Baseline + RCB + FF + DUC	0.650	0.766

Table 2. Leader board Dice scores of 2017 Liver Tumor Segmentation (LiTS) Challenge

Method	Per case	Global
H-DenseUNet [6]	0.722	0.824
Res-UNet [3]	0.670	-
Cascaded ResNet [15]	0.640	-
3D AH-Net [16]	0.634	0.834
FCN+Detector [17]	0.590	-
ours	0.650	0.766

4. CONCLUSIONS

In this work, we proposed a Feature-fusion Encoder-Decoder Network (FED-Net) based 2D deep learning model for liver lesion segmentation. Our feature fusion method uses the attention mechanism. By assigning different weights to different features according to the contribution of the feature to the final segmentation result, more useful features can be extracted. Our feature fusion method can effectively embed more semantic information into low-level features. We also propose a dense upsampling convolution instead of some conventional upsampling operations to compensate for the missing details during the upsampling process and add residual convolutional blocks to refine the target boundary information. Our approach demonstrates very promising performance relative to 2.5D and 3D models, and our 2D model is lightweight and can be easily deployed. In addition, our network has good generalization capabilities and can be applied to other areas of segmentation tasks, not just medical images.

5. REFERENCES

- [1] Patrick Ferdinand Christ, Mohamed Ezzeldin A. Elshaer, Florian Ettlinger, Sunil Tatavarty, Marc Bickel, Patrick Bilic, Markus Rempfler, Marco Armbruster, Felix Hofmann, Melvin DAnastasi, Wieland H. Sommer, Seyed-Ahmad Ahmadi, and Bjoern H. Menze, "Automatic liver and lesion segmentation in ct using cascaded fully convolutional neural networks and 3d conditional random fields," in *MICCAI*, 2016, pp. 415–423.
- [2] Avi Ben-Cohen, Idit Diamant, Eyal Klang, Michal Amitai, and Hayit Greenspan, *Fully Convolutional Network for Liver Segmentation and Lesions Detection*, Springer International Publishing, 2016.
- [3] Xiao Han, "Automatic liver lesion segmentation using a deep convolutional neural network method," *arXiv preprint arXiv:1704.07239*, 2017.
- [4] Qikui Zhu, Bo Du, Baris Turkbey, Peter Choyke, and Pingkun Yan, "Exploiting Interslice Correlation for MRI Prostate Image Segmentation, from Recursive Neural Networks Aspect," *Complexity*, vol. 2018, pp. 10, 2018.
- [5] Q. Zhu, B. Du, J. Wu, and P. Yan, "A deep learning health data analysis approach: Automatic 3d prostate mr segmentation with densely-connected volumetric convnets," in *2018 International Joint Conference on Neural Networks (IJCNN)*, July 2018, pp. 1–6.
- [6] Xiaomeng Li, Hao Chen, Xiaojuan Qi, Dou Qi, Chi Wing Fu, and Pheng Ann Heng, "H-denseunet: Hybrid densely connected unet for liver and tumor segmentation from ct volumes," *IEEE Transactions on Medical Imaging*, vol. PP, no. 99, pp. 1–1, 2017.
- [7] Qi Dou, Hao Chen, Yueming Jin, Lequan Yu, Jing Qin, and Pheng-Ann Heng, "3d deeply supervised network for automatic liver segmentation from ct volumes," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2016, pp. 149–157.
- [8] Olaf Ronneberger, Philipp Fischer, and Thomas 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.
- [9] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5168–5177.
- [10] Zhenli Zhang, Xiangyu Zhang, Chao Peng, Dazhi Cheng, and Jian Sun, "Exfuse: Enhancing feature fusion for semantic segmentation," *arXiv preprint arXiv:1804.03821*, 2018.
- [11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [12] Matthew D Zeiler and Rob Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*. Springer, 2014, pp. 818–833.
- [13] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," *arXiv preprint arXiv:1709.01507*, vol. 7, 2017.
- [14] Wenzhe Shi, Jose Caballero, Ferenc Huszar, Johannes Totz, Andrew P. Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Computer Vision and Pattern Recognition*, 2016, pp. 1874–1883.
- [15] Lei Bi, Jinman Kim, Ashnil Kumar, and Dagan Feng, "Automatic liver lesion detection using cascaded deep residual networks," *arXiv preprint arXiv:1704.02703*, 2017.
- [16] Siqi Liu, Daguang Xu, S Kevin Zhou, Olivier Pauly, Sasa Grbic, Thomas Mertelmeier, Julia Wicklein, Anna Jerebko, Weidong Cai, and Dorin Comaniciu, "3d anisotropic hybrid network: Transferring convolutional features from 2d images to 3d anisotropic volumes," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 851–858.
- [17] Míriam Bellver, Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Xavier Giró-i Nieto, Jordi Torres, and Luc Van Gool, "Detection-aided liver lesion segmentation using deep learning," *arXiv preprint arXiv:1711.11069*, 2017.