Fully convolutional network with sparse feature-maps composition for automatic lung tumor segmentation from PET images

Tian, Haihong, Xiang, Dehui, Zhu, Weifang, Shi, Fei, Chen, Xinjian

Haihong Tian, Dehui Xiang, Weifang Zhu, Fei Shi, Xinjian Chen, "Fully convolutional network with sparse feature-maps composition for automatic lung tumor segmentation from PET images," Proc. SPIE 11313, Medical Imaging 2020: Image Processing, 113131O (10 March 2020); doi: 10.1117/12.2548670

Event: SPIE Medical Imaging, 2020, Houston, Texas, United States

Fully Convolutional Network with Sparse Feature-maps Composition for Automatic Lung Tumor Segmentation from PET Images

Haihong Tian[a], Dehui Xiang[a], Weifang Zhu[a], Fei Shi[a], Xinjian Chen[a]
[a]School of Electronics and Information Engineering, Soochow University, Suzhou, Jiangsu Province,215006,China

ABSTRACT

Accurate lung tumor delineation plays an important role in radiotherapy treatment planning. Since the lung tumor has poor boundary in positron emission tomography (PET) images, it is a challenging task to accurately segment lung tumor. In addition, the heart, liver, bones and other tissues generally have the similar gray value as the lung tumor, therefore the segmentation results usually have high false positive. In this paper, we propose a novel and efficient fully convolutional network with a trainable compressed sensing module and deep supervision mechanism with sparse constraints to comprehensively address these challenges; and we call it fully convolutional network with sparse feature-maps composition (SFC-FCN). Our SFC-FCN is able to conduct end-to-end learning and inference, compress redundant features within channels and extract key uncorrelated features. In addition, we use deep a supervision mechanism with sparse constraints to guide the features extraction by a compressed sensing module. The mechanism is developed by driving an objective function that directly guides the training of both lower and upper layers in the network. We have achieved more accurate segmentation results than that of state-of-the-art approaches with a much faster speed and much fewer parameters.

Keywords: Lung Tumor Segmentation, Fully Convolutional Network, Sparse Feature-maps Composition

1. INTRODUCTION

Lung cancer is the most common cause of cancer-related death worldwide in both men and women [1]. Automatic tumor segmentation is important for many clinical applications, such as treatment effect measuring, radiation treatment planning, and robust features extraction for high-throughput radiomics. However, it is still a highly challenging task due to the complex background, fuzzy boundary, and irregular shape of the lung tumor in medical images. In addition, the heart, liver, bones and other tissues generally have the similar gray value as the lung tumor, and therefore the segmentation results usually have high false positive.

PET images have been commonly used for tumor delineation in clinical radiotherapy applications due to their high contrast to non-tumor tissues. Several methods have been proposed for automatic lung tumor segmentation from PET images. Generally, these methods can be categorized into non-learning-based and learning-based approaches. Non-learning-based methods usually rely on the statistical distribution of the intensity, including SUV thresholding, clustering, and graph-based methods [2]. However, these methods usually have too limited representation capability to deal with the large variations of tumor shape. On the other hand, learning-based approaches take the advantage of hand-crafted features to train classifiers to achieve good segmentation.

Recently, deep learning has been shown to achieve superior performance in various challenging tasks, such as classification, segmentation, and detection. Deep learning has quickly proved to be the most advanced tool for dealing with various medical image processing tasks, including segmentation. The original FCN was proposed in [3], and many variants have been developed in the field of medical image processing, including UNet [4], VNet [5] and etc. They have shown remarkable success in a variety of computer vision and medical image.

*Corresponding author: Dehui Xiang, E-mail: xiangdehui@suda.edu.cn

Medical Imaging 2020: Image Processing, edited by Ivana Išgum, Bennett A. Landman,
Proc. of SPIE Vol. 11313, 113131O · © 2020 SPIE · CCC
code: 1605-7422/20/$21 · doi: 10.1117/12.2548670

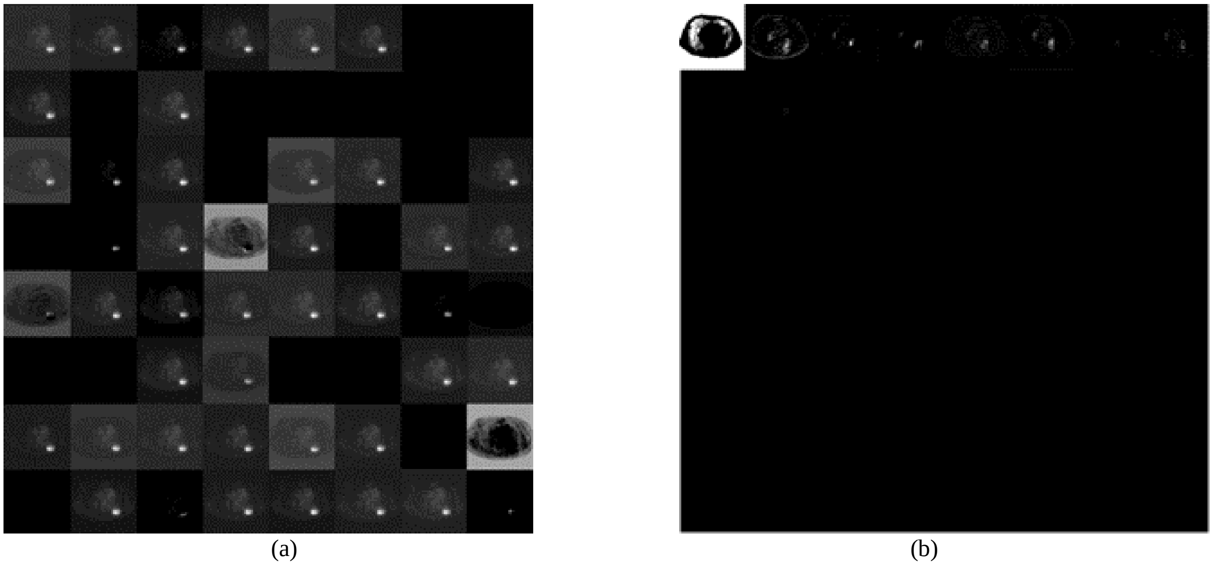

Fig. 1. The middle layer feature maps from a well-trained network. (a) is a middle layer feature maps from UNet, the output size is 256×256×64. (b) is a middle layer feature maps from our proposed network, the output size is 256×256×64

However, CNN performs poor segmentation when the target has the highly similar intensities with its surrounding structures especially in lung tumor segmentation. Due to the high intensities of tumors in PET images, other organs such as the heart, spine and liver have similar intensities to tumors. This has led to high false positives when networks such as UNet are used. First, most of the CNN are unable to discriminate which feature map is more effective to the result in each layer. The feature maps are not discriminative to differentiate. Therefore, it is important to extract and select key features of lung tumors and remove similar responses from surroundings. Second, the number of filters for each layer needs to be manually tried many times when CNN is used for segmentation tasks. In order to capture all the features, we usually set the number of the filters to be large, which may result in the oblivion of key features. There is a great need to automatically determine the number of filters. As shown in Fig.1, in one the middle layer of the trained UNet, it can be seen that many feature channels have the similar feature maps, which results in information redundancy. Therefore, it is important to preserve and enhance the effective features of lung tumors and choose the appropriate parameters of CNN.

Principal component analysis (PCA) was invented by Pearson [6]. As a dimension reduction and feature extraction method, PCA has numerous applications in statistical learning, such as handwritten zip code classification, human face recognition, eigengenes analysis, gene shaving. Such dimensionality reduction can be a very useful step for visualizing and processing high-dimensional datasets, it can reduce different possible explanatory variables to a few principal components while still retaining as much of the variance in the dataset as possible. Therefore, in our proposed network, we define a trainable compressed sensing module with PCA to preserve and enhance the effective features of lung tumors and remove similar responses from surroundings. In addition, a deep supervision mechanism with sparse constraints is also proposed to comprehensively address these challenges.

Our contributions are summarized as follows:

- A trainable compressed sensing module called CSM is proposed. It can implement information compression, remove redundant feature maps, and enhance effective feature maps during the training procedure.
- Our proposed network can extract key features by CSM and increase these key features by convolutional layers so as to obtain excellent segmentation results.
- A deep supervision mechanism is proposed to supervise the weights in CSM. Our deep supervision mechanism guides the features extraction with CSM. Such a mechanism is developed by driving an objective function that directly guides the training of both lower and upper layers in the network.
- Our network can achieve more accurate segmentation results than that of state-of-the-art approaches with a much faster speed and much fewer parameter

2. METHODS

Fig.2 shows the whole pipeline of our proposed fully convolutional network with sparse feature-maps composition (SFC-FCN). The input is PET images, and the out-put is a binary segmentation of lesions. The SFC-FCN is first trained by using the images with manually annotated lesions to extract features. In the testing process, the images that unknown to SFC-FCN is used to generate predictions. The details of our proposed method are introduced in the following subsections.

2.1 Data Augmentation

Note that each patient has a limited number of lesions, and therefore there are only a small number of slices with lung tumors in the whole 3D images. It is thus necessary to generate a large number of slices to tune the massive network parameters. To balance the number of the normal and lesion slices, we only perform data augmentation on lesion slices by horizontally flipping, shifting and randomly rotating the slices.

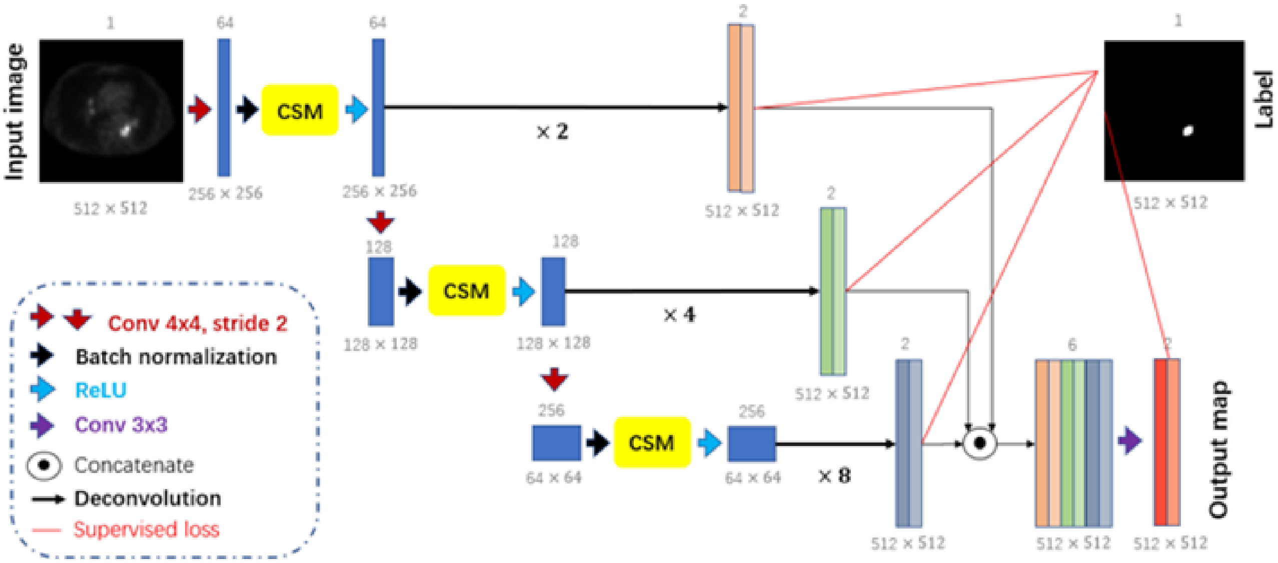

Fig.2. An overview of the sparse feature-maps composition fully convolutional network (SFC-FCN). The yellow CSM is the compressed sensing module.

2.2 Network Architecture

Fig.2 shows the network structure of our proposed method. In this paper, we propose to use fully convolutional network (FCN) as the base network for key feature extraction with our compressed sensing module (CSM), and deconvolution from deep supervision layers for score map. The details of the CSM and deep supervision mechanism are shown in section 2.3 and 2.4.

2.3 Compressed Sensing Module

PCA can be used as a method of compressed sensing. Eigen decomposition is the most often use to solute a PCA problem. The Eigen decomposition formulation of PCA also relates PCA to the singular value decomposition (SVD). Since SVD can be interpreted as the best low-rank approximation to the data matrix, we perform SVD as the solution to the compressed sensing.

As illustrated in Fig.3, the input of the CSM is the output from last layer. Let the input of CSM be$X \in \mathbb{R}^{H \times W \times C}$, where H is the height of feature maps, W is the width of feature maps and C is the channels of current layer. Then we reshape them to $\mathbb{R}^{N \times C}$, where $N = H \times W$ is the number of pixels in one feature map. We perform SVD on the reshaped matrix X.

Matrix X can be factorized as $\mathrm{X} = \mathrm{U}\Sigma\mathrm{V}^{\mathrm{T}}$, where $\mathrm{U} \in \mathbb{R}^{N \times N}$ and $\mathrm{V} \in \mathbb{R}^{C \times C}$ are both orthogonal matrices and $\Sigma \in \mathbb{R}^{N \times C}$ is a matrix whose elements are nonnegative real numbers on the diagonal and zero elsewhere. The diagonal elements $\{\lambda_i\}_{i=1}^{C}$, referred to as singular values, are sorted in descending order, and the columns of U and V are referred to, respectively, as left and right singular vectors. The covariance matrix may be written as:

$$XX^T = \mathrm{U}\Sigma\mathrm{V}^\mathrm{T} V\Sigma^\mathrm{T}\mathrm{U}^\mathrm{T} = \mathrm{U}\Sigma\Sigma^\mathrm{T}\mathrm{U}^\mathrm{T} \quad (1)$$

where $\Sigma\Sigma^\mathrm{T}$ is a diagonal matrix. Thus, singular-value decomposition of X is equivalent to eigen decomposition of XX^T. The singular values of X are the square roots of the eigenvalues of XX^T, and the left singular vectors of X are the eigenvectors of XX^T. Eigen decomposition is unique up to the scale of the eigenvectors, which we normalize, and to permutations of the eigenvectors and their corresponding eigenvalues, which we sort in descending order.

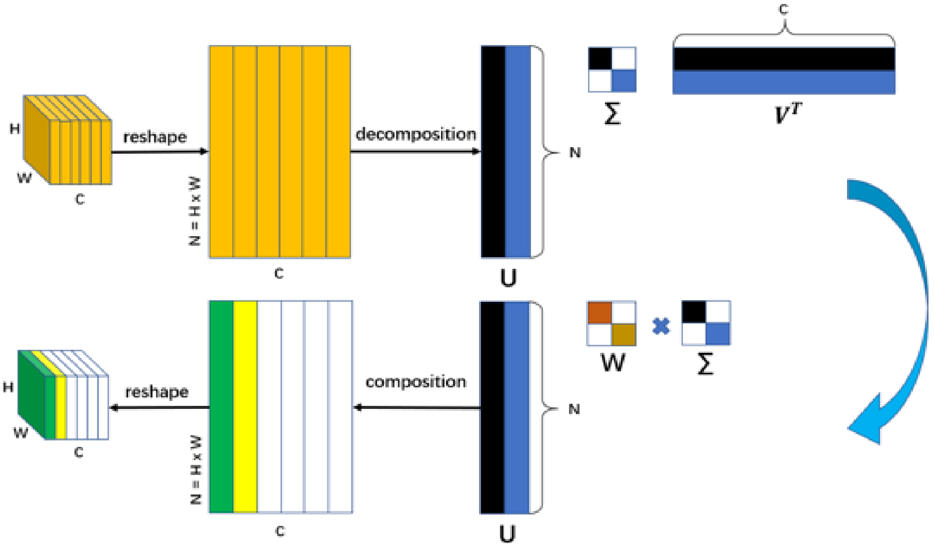

Fig. 3. The proposed compressed sensing module (CSM). The orange block represents equal feature maps input. The output in green, yellow and white colors represents sparse feature-maps composition.

Therefore, if we want to reduce dimension, we can follow: $Y_r = U\Sigma_r$, where Y_r is the matrix after SVD dimensionality reduction, r is the first largest singular values. However, it is not trainable in this way of information compression. Since that the singular values are sorted in descending order, we can define a weight matrix φ, whose elements are trainable parameters on the diagonal and zero elsewhere, has the same shape as matrix Σ. Our new Y can be defined as:

$$\mathrm{Y} = \mathrm{U}\,(\varphi\Sigma) \quad (2)$$

With the trainable matrix φ, the CSM can implement information compression, remove redundant information, and enhance effective information during the training procedure. As a result, the number of channels in the network remains the same, but the total amount of information is reduced in each layer, so the feature maps of the layer become sparse and uncorrelated with each other.

Each CSM can be understood as an information compressing technique to convert correlated feature maps into a set of linearly uncorrelated feature maps called principal components. This is consistent with our goal of obtain uncorrelated feature maps. Finally, we can obtain a set of sparsely feature maps by the CSM.

2.4 Deep Supervision Mechanism

We use deep a supervision mechanism with sparse constraints to guide the features extracted by CSM.

Specifically, we first up-scale some lower-level and middle-level feature volumes using additional deconvolutional layers. Then, we employ the softmax function on these full-sized feature volumes and obtain extra dense predictions. For these branched prediction results, we calculate their classification errors (i.e., softmax-crossentropy) with regard to the ground truth segmentation masks. These auxiliary losses together with the loss from the last output layer are integrated to energize the back-propagation of gradients for more effective parameter updating in each iteration.

Let φ^l be the weights of the l th (l = 1, 2, …, L) CSM in our network, we denote the set of all CSM weights by Φ = (φ^1, φ^2, …, φ^L). With $\mathrm{p}(x_i;\ \Phi)$ representing the probability prediction of a pixel x_i after the softmax function in the last output layer, the crossentropy loss can be formulated as:

$$\mathcal{L}_{ce}(\ \mathcal{X};\ \Phi) = -\sum_{x_i \in \mathcal{X}} y_i \log \mathrm{p}(x_i;\ \Phi) + \lambda \|\Phi\|^2 \tag{3}$$

where $\mathcal{X}$ represents the training dataset and y_i is the target class label corresponding to the pixel $x_i \in \mathcal{X}$, the second term is the weight regularization and λ is the trade-off hyperparameter.

On the other hand, the deep supervision is exactly introduced via branch networks as the red lines shown in Fig.2. To introduce deep supervision from the d th hidden layer, we denote the weights of the first d CSM in the mainstream network by Φ_d = (φ^1, φ^2, …, φ^d), and then the auxiliary loss for deep supervision can be formulated as:

$$\mathcal{L}_{ced}(\ \mathcal{X};\ \Phi_d) = -\sum_{x_i \in \mathcal{X}} y_i \log \mathrm{p}(x_i;\ \Phi_d) \tag{4}$$

We can learn the best weights Φ, so as to supervise our CSM to extract key effective feature maps.

2.5 Loss Function

Although we use data augmentation to enlarge the number of slices which have lesions, the data imbalance is still present in a single slice because our segmentation object tumor is too small and only occupies a small part of the slice. In order to remedy this imbalance, we use focal loss [7] to supervise the output of last layer, as shown in Equation (5). We adjust the parameters α and γ of focal loss for better performance.

$$\mathcal{L}_{focal}(\ \mathcal{X}\ ) = -\sum_{x_i \in \mathcal{X}} \alpha (1 - p_k)^{\gamma} \log p_k \tag{5}$$

where $p_k = y_i \mathrm{p}(x_i)$, is the estimated probability for class k.

The total loss consists of three parts: auxiliary softmax-crossentropy loss from middle layers, and softmax-crossentropy loss, focal loss from the last layer. The total loss for our network can be seen in Equation (6).

$$\mathcal{L}_{total}(\ \mathcal{X};\ \Phi) = \mathcal{L}_{ce}(\ \mathcal{X};\ \Phi) + \sum_{\mathrm{d} \in \mathrm{D}} \eta_d \mathcal{L}_{ced}(\ \mathcal{X};\ \Phi_d) + \delta \mathcal{L}_{focal}(\ \mathcal{X}) \tag{5}$$

where η_d is the balance weight of $\mathcal{L}_{ced}$, and δ is the balance weight of $\mathcal{L}_{focal}$, and D is the set of index of all the hidden layers which are equipped with the deep supervision.

3. RESULTS

3.1 Datasets

Our segmentation approach was evaluated in a data set which consists of 54 3D PET images obtained from different patients with non-small cell lung cancer (NSCLC). All the PET image size is 512 × 512 × 60, the voxel size is 0.234 × 0.234 × $1mm^3$. The reference segmentation was obtained by two professional clinical experts manually on the PET images by the guidance of the corresponding CT images.

3.2 Data Prepossess and Data Augmentation

In our experiments, we divided 54 3D PET images into 49 training images and 5 testing images, and then cut them into 2D slices for training. However, taking the imbalance data distributing into consideration, we need to do data augmentation using flip, rotation and width or height shift on slices which have lesions. We get a balanced training set, and the total nearly training number is 14700. 13-fold training and testing were conducted.

3.3 Implementation Details

For data augmentation, we used horizontal flip, rotation and width/height shift with keras ImageDataGenerator to triple each lesion image. The corresponding scales are {True, 0.2, 0.2}. We trained the network using stochastic gradient descent (SGD) with batch size 4, momentum 0.9 and weight decay 0.0001. We used the "poly" learning rate policy where the learning rate is multiplied by $\left(1 - \frac{iter}{\max\ \_iter}\right)^{power}$ with power 0.9 and initial leaning rate $4e^{-3}$.

3.4 Segmentation Results

To quantitatively assess the performance of our proposed method, we compared the segmentation results with the ground truth according to the following four metrics: dice similarity coefficient (DSC), precision, true positive fraction (TPF) and false positive fraction (FPF). The DSC calculates the overlap between the segmentation results and ground truth, and is defined as:

$$\text{DSC} = \frac{2TP}{FP+2TP+FN} \tag{6}$$

where, TP is the number of true positives, FP is the number of false positives and FN is the number of false negatives. TPF and precision metrics are computed as:

$$\text{TPF} = \frac{TP}{TP+FN} \tag{7}$$

$$\text{Precision} = \frac{TP}{FP+TP} \tag{8}$$

We compared the segmentation results to previous methods as shown in Table 1. Some segmentation results of our method was shown in Fig.4. Some segmentation results of different methods were shown in Fig.5. In order to prove the validity of the proposed module, the corresponding ablation segmentation results were as shown in Table2. Method baseline was based on simplified FCN-8.

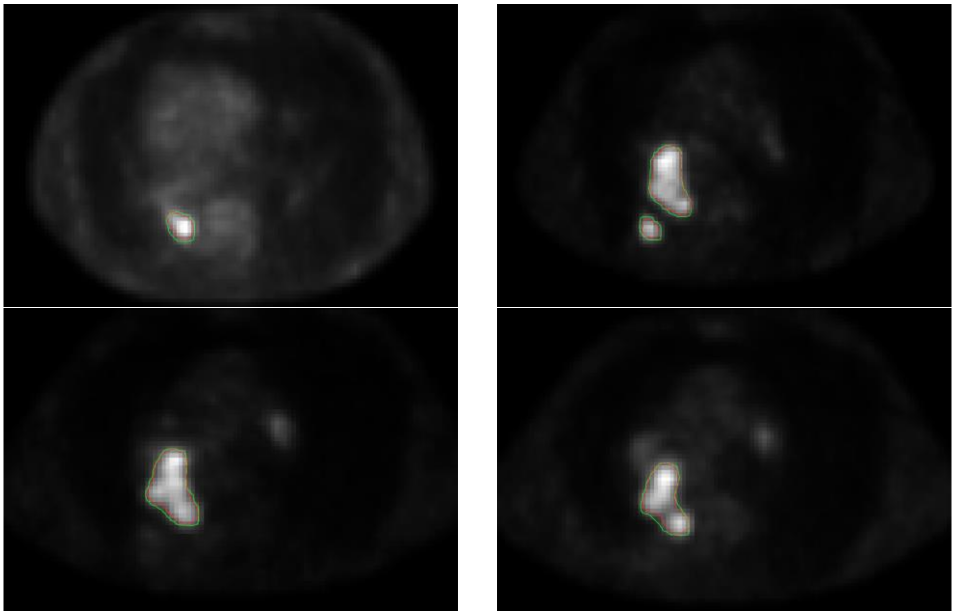

Fig. 4. Segmentation results our method. The red curve represents our segmentation results. The green curve represents the ground truth.

Table 1. Comparison of the quantitative segmentation results for different methods (mean±standard deviation).

Methods	DSC (%)	Precision (%)	TPF (%)	FPF (%)	Param
DenseNet [8]	57.77±17.31	57.59±26.55	87.47±15.65	0.29±0.24	2.4M
CGAN[9]	54.10±10.01	75.40±14.58	51.57±16.08	5.66±13.39	142M
Segcaps[10]	58.98±23.21	50.03±23.43	**97.89±3.01**	3.29±3.46	1.4M
UNet	58.04±18.56	60.10±20.08	90.65±10.71	0.12±0.11	31M
SFC-FCN	**79.63±7.99**	**86.83±7.14**	92.05±5.81	**0.02±0.01**	**1M**

Table 2. The corresponding ablation segmentation results (mean±standard deviation).

Methods	DSC (%)	Precision (%)	TPF (%)	FPF (%)
Baseline	46.73±20.98	57.59±26.55	50.52±21.78	0.16±0.15
Baseline with deep supervision	59.79±21.58	76.10±15.71	66.95±13.88	0.14±0.11
Baseline with CSM	66.42±24.65	83.31±7.39	74.89±3.01	0.02±0.08
Our SFC-FCN	**79.63±7.99**	**86.83±7.14**	**92.05±5.81**	**0.02±0.01**

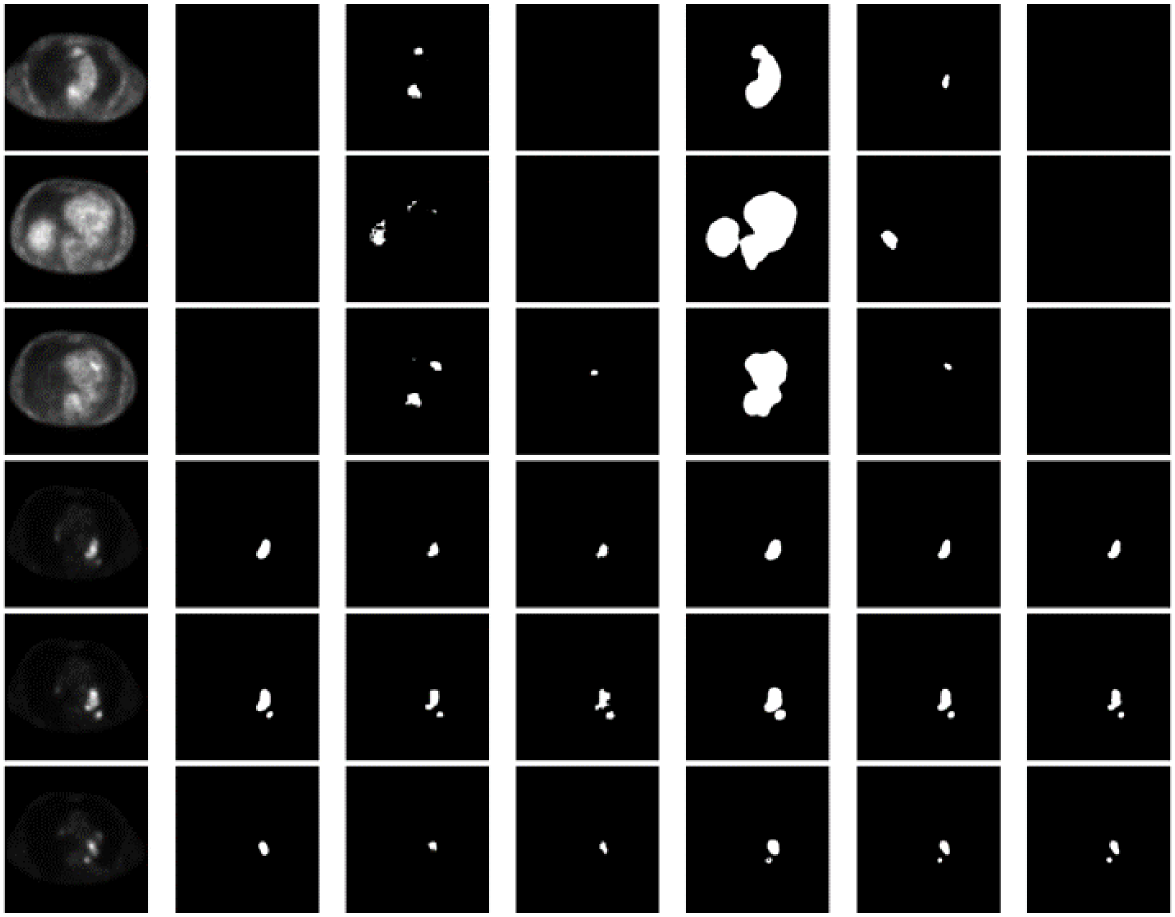

Fig. 4. Segmentation results of different methods. The first column is the original image; the second column is the ground truth; the next columns are the results of CGAN, DenseNet, Segcaps, UNet; the last column is our segmentation results.

4. CONCLUSIONS

In this paper, we proposed an automatic lung tumor segmentation algorithm based on fully convolutional network with a trainable compressed sensing module and deep supervision mechanism. Our proposed network could extract key uncorrelated features by CSM and increased these key features by convolutional layers so as to obtain excellent segmentation results and could achieve competitive segmentation results to state-of-the-art approaches with a much faster speed and much fewer parameters.

5. ACKNOWLEDGEMENTS

This work has been supported in part by the National Key R&D Program of China under Grant 2018YFA0701700, and in part by the National Natural Science Foundation of China (NSFC) under Grant 61971298.

6. REFERENCE

[1] Stewart, B. W. K. P., and Christopher P. Wild. "World cancer report 2014," Public Health, (2014).

[2] Ju, Wei, et al. "Random walk and graph cut for co-segmentation of lung tumor on PET-CT images," IEEE Transactions on Image Processing, 24(12), 5854-5867, (2015).

[3] Evan Shelhamer, Jonathan Long, Trevor Darrell, “Fully Convolutional Networks for Semantic Segmentation,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(4), 640-651, (2017).

[4] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in Proceedings of International Conference on Medical Image Computing & Computer-assisted Intervention, 234–241, (2015).

[5] F. Milletari, N. Navab, and S.-A. Ahmadi,” V-net: Fully convolutional neural networks for volumetric medical image segmentation,” in Proceedings 4th International Conference 3D Visualization, 565–571, (2016).

[6] K. Pearson, “Liii. on lines and planes of closest fit to systems of points in space,” Philos. Mag. J. Sci., 2(11), 559–572, (1901).

[7] Lin T Y, Goyal P, Girshick R, et al. “Focal Loss for Dense Object Detection,” Proceedings of the IEEE international conference on computer vision, 2980-2988, (2017).

[8] Huang G, Liu Z, Maaten L V D, et al. “Densely Connected Convolutional Networks,” 2017 IEEE Conference on Computer Vision and Pattern Recognition, IEEE Computer Society, 4700-4708, (2017).

[9] Isola P, Zhu J Y, Zhou T, et al. “Image-to-Image Translation with Conditional Adversarial Networks,” Proceedings of the IEEE conference on computer vision and pattern recognition, 1125-1134, (2017).

[10] Lalonde R, Bagci U. "Capsules for object segmentation," arXiv preprint arXiv:1804,04241, (2018).