3D ENHANCED MULTI-SCALE NETWORK FOR THORACIC ORGANS SEGMENTATION

Qin Wang[1,2] *, Weibing Zhao*[1,2] *, Chunhui Zhang*[1,2] *, Liyue Zhang*[3]*, Changmiao Wang*[1,2]*, Zhen Li*[⋆ 1,2]
Shuguang Cui[1,2] , Guanbin Li[3]

[1] The Chinese University of HongKong (Shenzhen), China
[2] Shenzhen Research Institute of Big Data, China
[3]Sun Yat-Sen University, China

ABSTRACT

We focus on the segmentation of 4 OAR: heart, aorta, trachea, esophagus through Computed Tomography (CT) thoracic scans, where the esophagus is the most challenging to segment due to its volume imbalance and narrow shape. In this paper, a 3D Enhanced Multi-scale Network (EMSN) is proposed to improve the performance, which is a refined variant of the 3D FCN network. Specifically, an extra stage is added to refine the prediction map by concatenating a preliminary prediction map with the CT image to utilize auto-context. 3D dilated convolution is employed to enlarge the receptive field of convolution kernel without loss of resolution. Besides, more residual connections are added in V-Net to avoid gradient degradation during back-propagation. For data preprocessing, a maximum bounding box is calculated and used to crop raw data to reduce calculation cost. And to achieve data augmentation, registration is performed by aligning each CT image and corresponding ground truth with the others in training set to generate non-linearly transformed new CT images along with annotations. The consensus is utilized by retaining models in the last few epochs to segment and vote for the final prediction. Experiments demonstrate that our EMSN model achieves competitive performance on SegTHOR dataset.

***Index Terms*—** CT Segmentation, V-Net, multi-scale refine, auto-context, image registration

1. INTRODUCTION

Tumor disease is the most common cause of disease deaths. Using screening with low-dose computed tomography (CT) is essential for this situation [1]. The target tumor and healthy organs located near the target tumor is called Organs at Risk (OAR) which need to be delineated for further irradiation planning.

Some studies have aimed to OAR problem in automatic segmentation over the years. Recently, deep learning-based multi-organ segmentation in abdominal CT images has been approached. Deep learning methods have achieved state-of-the-art performance due to the strong non-linear modeling capability. For U-Net [2] in 2D medical image segmentation, skip connections between upsampling and downsampling layers combine high-resolution features with the upsampled output. Then Milletari developed U-Net to 3D V-Net [3]. Both of them inspired many new frameworks extensively [4]. For example, the DenseV-Net segmentation network is proposed to enable high-resolution activation maps due to feature reuse, and it achieves good accuracy on 3D CT images [5]. Multi-scale pyramid 3D-FCNs (MSN) perform good learning properties for end-to-end training, and it improves segmentation accuracy of fine structures [6].

However, it's still challenging to segment organs with narrow structures like the esophagus, whose volume is also explicitly unbalanced compared with other organs like the heart. Inspired by the multi-scale pyramid of 3D FCN [6], we propose a 3D Enhanced Multi-Scale Network to overcome the difficulties mentioned above, where the third stage is employed to enhance performance. In the last two stages, contextual information is fused with a high-resolution image as deep supervision. Preliminary prediction maps are aggregated with image features to improve the overall segmentation accuracy. Our proposed EMSN is trained in an end-to-end fashion, which achieves a promising segmentation result on the SegTHOR dataset.

In summary, the key contributions are 3-fold: (1) We enhance the Multi-scale network by adding one more stage network and increasing the number of parameters on each network, as well as dilated convolution layers employed in EMSN. (2) According to the maximum bounding box, the preprocessing method is used to crop the raw data to reduce heavy calculation burden and improve the accuracy. (3) We greatly enlarge the training dataset by using the Nifty image registration software to change our training data from 40 to 1600, for faster convergence and higher accuracy, and we use the weights of pretrained abdomen dataset to initialize our model.

2. METHODS

Holger *et al.* [6] proposed a multi-scale pyramid of 3D FCN network as shown in Figure 2. As shown in Figure 1, we propose an Enhanced Multi-scale Network (EMSN) by adding a third stage and other boosting methods, which will be illustrated as follows.

2.1. Enchanced Multi-scale Networks

In the first stage, CT scan is downsampled and then fed into V-Net to train a coarse segmentation map in low resolution to delineate an approximate location of organs. Basically, lower resolution downsampled prediction maps have more contextual information, while high-resolution images are aimed for local accurate segmentation. So the coarse segmentation map is upsampled to the original size and then concatenated with the original input CT image to aggregate multi-scale contextual information. Assisted by the coarse prediction map, V-Net in stage 2 outputs a better segmentation result.

We propose to add one more stage to refine segmentation further. Since scale based auto-context improves the segmentation performance to a large extent, the segmentation map from stage 2 is concatenated with the original input cuboid again to refine the prediction map. The architecture is illustrated in Figure 1.

Denote $\mathcal{V}(\cdot)$ as the operation of V-Net which can segment an input of 3D image $\mathcal{X}$ to a segmentation map $\mathcal{S}$, i.e., $\mathcal{S} = \mathcal{V}(\mathcal{X})$. Denote $\oplus$ as a concatenating operation. The superscript ds and us indicate downsampling and upsampling operation. Then the process of our model can be illustrated as,

$$\mathcal{S}_1 = \mathcal{V}(\mathcal{X}^{\text{ds}}) \tag{1}$$

$$\mathcal{S}_2 = \mathcal{V}(\mathcal{X} \oplus \mathcal{S}_1^{\text{us}}) \tag{2}$$

$$\mathcal{S}_3 = \mathcal{V}(\mathcal{X} \oplus \mathcal{S}_2) \tag{3}$$

$$\hat{\Theta} = \arg\min_{\Theta} \frac{1}{3} \sum_{i=1}^{3} \mathcal{L}(\mathcal{S}_i, \Theta, L) \tag{4}$$

where Θ denotes the network parameters, such as convolutional kernel weights. And L is ground truth label image. $\mathcal{L}$ is the loss function which will be introduced in 2.2.

2.2. Loss function

The net makes voxel-to-voxel predictions. Like the binary case in [3], it can be trained by minimizing a dice loss function:

$$\mathcal{L}_j\left(\mathcal{X}_j, \Theta_j, L_j\right) = 1 - \frac{1}{K} \sum_{k=1}^{K} \left(\frac{2 \sum_i^N p_{i,k} \cdot l_{i,k}}{\sum_i^N p_{i,k}^2 + \sum_i^N l_{i,k}^2} \right). \tag{5}$$

where $p_{i,k} \in [0, \ldots, 1]$ represents the continuous values of the *softmax* 3D prediction maps for each class label k of K, $l_{i,k}$ represents the corresponding ground truth.

The final loss function joints 3 stages' loss functions together, as following

$$\mathcal{L}\left(\mathcal{X}, \Theta, L\right) = \sum_{j=1}^{3} \left(\mathcal{L}_j\left(\mathcal{X}_j, \Theta_j, L_j\right)\right). \tag{6}$$

where j denotes the order number of the stage network in the EMSN. We train the 3-stage network end to end with the joint loss function 6.

2.3. Residual V-net

V-Net [3] is a volumetric convolution neural network which performs segmentation on CT images as shown in Figure 1. Residual connections are applied to our enhanced model to avoid gradient vanishing problem. Encoder and decoder both contain 4 blocks. This architecture performs voxel-to-voxel predictions [7]. For the encoder, the first block has only one convolution layer with kernel size 3, and the second block has 2 layers. The third and fourth blocks both have 4 layers, and each layer has the dilation with size 4. And there are downscale convolution layers with stride 2 to downsample the size between two adjacent blocks. The number of output channels of the downscale convolution layer is two times the number of input channels. The decoder architecture is similar to the inverse procedure of encoder, except using transpose convolution layer to upscale the feature maps instead of downscale convolution layers in the encoder. To enhance the fitting ability and avoid the over-fitting problem, the dropout rate [8] is set to be 0.3. Skip connections between encoder and decoder are similar to U-Net [2] to get a better optimization result and faster convergence.

2.4. Increase the number of Parameters

Parameters are increased in order to capture more features of data and enhance the fitting ability of the model. Taking the GPU memory limitation into account, we just increase the deeper blocks in encoder and decoder path, because the deeper block's feature maps are in low resolutions relatively. So we double the number of layers on the two deepest blocks from 3 to 6 layers, and break 6 layers into two parts, with 3 layers in each part. Two parts are connected by a residual connection to avoid gradient vanish. Besides, we add a new stage 3 network which can also be considered as increasing the number of parameters.

2.5. 3D Dilated Convolution

In the compression path of V-Net, downsampling is performed after each block to enlarge the receptive field, while in the decompression path, a de-convolution operation is employed to recover the size of the feature map to realize pixel-wise segmentation. Internal data structures and spatial

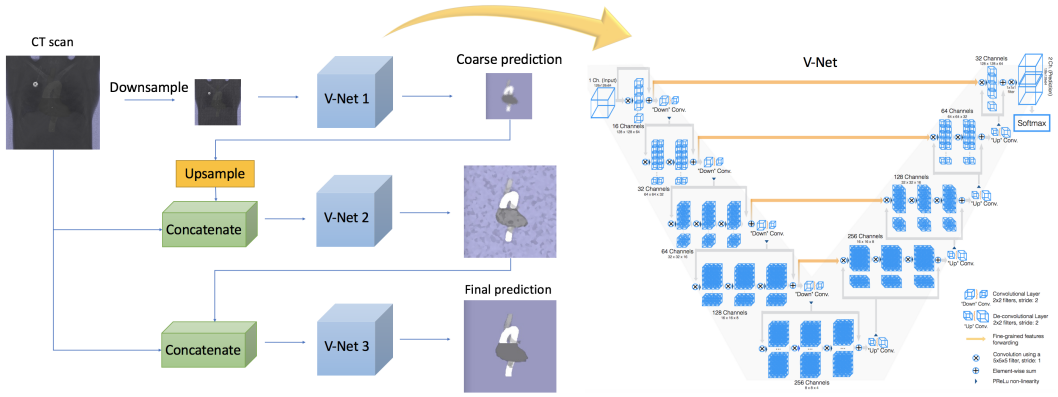

Fig. 1. Enhanced Multi-scale Network for thoracic organs segmentation in CT images. In the first level, CT scan is downsampled to a low resolution and then fed into V-Net 1 to predict a rough segmentation map, which will be aggregated with original CT image as a deep supervision. Therefore V-Net 2 can learn the contextual information and image feature information simultaneously to provide a more accurate segmentation map in high resolution, which will be concatenated with the input image again in level 3 to further refine segmentation result.

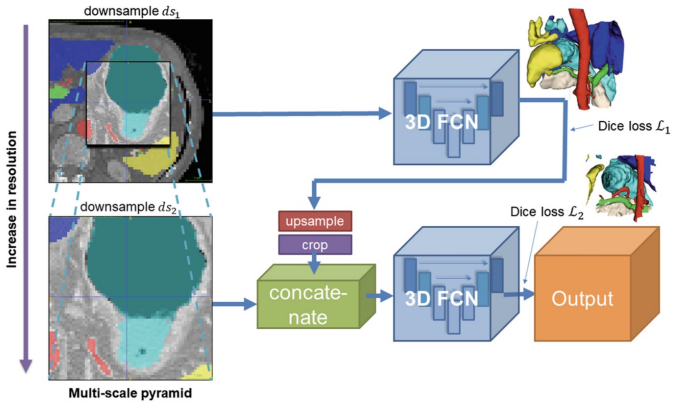

Fig. 2. Multi-scale Pyramid 3D FCN Network

hierarchical information is lost during this series of downsampling and upsampling. So 3D dilated convolution [9] is employed to enlarge the receptive field with the size of the feature map remains unchanged, which can maintain the resolution and coverage.

2.6. Consensus

It's biased for the model to do the segmentation task which loads only one weights file at the end epoch of training. In order to avoid errors and improve accuracy, we preserve 5 weights of the model in the very last time of the training progress. These different weighted models to segment CT image separately. Finally, we average the 5 probability maps voxel-wise and then assign each voxel to a specific class label with the largest probability. The prediction error is largely reduced by this multi-model consensus method.

3. EXPERIMENTS

Multi-scale Pyramid 3D FCN Network (MSN) as shown in Fig. 2 is chosen as the baseline for our experiment, which will be compared with our Enhanced Multi-scale Network (EMSN) to show the superiority of our proposed model.

3.1. Dataset

Abdomen Dataset [10] which consists of 13 classes of organs are utilized to pretrain our model. Since the pre-trained model has learned a rough outline about organs, the parameters of convolutional kernels are loaded from the shallow layers of the pre-trained model to initialize our model for a better prediction on the SegTHOR dataset.

SegTHOR dataset This dataset consists of 11084 slices from 60 patients and it has been randomly split into two parts: 40 patients, 7390 slices for the training set; 20 patients, 3694 slices for the testing set. The size of CT scans is 512×512 with a resolution of 0.90∼1.37 mm. The number of slices for every patient is different between 150 and 284. Besides, z-resolution varies from 0.90 mm to 1.37 mm. The patients have non-small cell lung cancer and they were fully anonymous. The reference standard for ground truths is manually annotated.

Abdomen dataset This dataset consists of 50 abdominal CT scans which are randomly selected from a combination of an ongoing colorectal cancer chemotherapy trial. The 50 scans were captured during portal venous contrast phase with variable volume sizes ($512 \times 512 \times 85 - 512 \times 512 \times 198$).

Thirteen abdominal organs were manually labeled and verified by a radiologist on a volumetric basis using the MIPAV software, including: (1) spleen (2) right kidney (3) left kidney (4) gallbladder (5) esophagus (6) liver (7) stomach (8) aorta (9) inferior vena cava (10) portal vein and splenic vein (11) pancreas (12) right adrenal gland (13) left adrenal gland.

3.2. Pre-processing

The internal organs of the human body are roughly in the same area. Therefore, we calculate a maximum bounding box through the dataset statistics, which guarantees that all the organs of interest in each sample in the dataset can be contained in the box. Based on the bounding box, raw data is cropped to get a dataset in a smaller size. After the cropping pre-processing, the size of each CT data becomes $304 * 224 * slices_{gt}$, where $slices_{gt}$ represents the number of slices in CT data containing ground truth organs. This cropping method can reduce the calculation cost and remove the noises to get better performance during the training.

Data augmentation In addition, we use data augmentation when preparing raw training data, rotation and scale in 0.7 probability are utilized. *NiftyReg* [11] is used to align training data (40 CT scans and ground truth) with all the other training data respectively. Thus, the training dataset can be enlarged to 40×39 CT images with annotations, which tremendously improves the scale of the training dataset.

3.3. Post-processing

The original prediction results have some noise like many small black clouds around the organs. In order to remove them, we use the largest connected component algorithm to save only the main area of the respective predicted organs of interest.

The resolution of prediction results should be resumed to raw dataset's resolution by padding zeros because we crop the datasets according to the bounding box during the pre-processing phase.

3.4. Training Details

The implementation of our networks is based on PyTorch-1.0.0. The maximum iteration for training is set to 3000 epochs. We use Adam as the optimizer with a learning rate of $1e-4$ which decays at 2000, 2500 epoch. We train our model end to end with joint loss in three stages on $4\times$ NVIDIA TITAN Xp(Pascal) GPUs. Due to the limitation of GPU memory, a patch of data with a size of $304 \times 224 \times 48$ is randomly selected from a raw dataset and then fed into the network. By this way, we train the model in patch based fashion.

3.5. Results

Dice metric(DM) [3] and Hausdorff distance(HD) [12] serve as the standard evaluation metrics. And the best segmentation results obtained from our experiments are shown in table 1. Especially, the organ which is hard to segment (e.g. esophagus) outperforms than other teams.

Individual Component Analysis Particularly, the segmentation performance of hard organs, like esophagus with long and narrow shape, is improved significantly by multi-scale refinement.

Comparison between different models As shown in Table 1, high performance on all organs is obtained by Multi-scale network(EMSN). EMSN* denotes training EMSN with data augmentation which includes initializing the weights by loading the Abdomen dataset pretrained model and around 1600 new samples are generated by NiftyReg. EMSN+ denotes increasing more convolution layers based on EMSN. EMSN*+ denotes including both two methods mentioned above. The experiments show that the best performance for each organ can be achieved by EMSN*+.

Table 1. Dice of Models on Different Organs

Organ \ Model	MSN	EMSN	EMSN*	EMSN*+
Esophagus	0.7664	0.8386	0.8294	**0.8597**
Heart	0.9159	0.9310	0.9432	**0.9459**
Trachea	0.8628	0.9067	0.9102	**0.9217**
Aorta	0.8907	0.9352	0.9341	**0.9433**
Mean	0.8589	0.9028	0.9042	**0.9176**

Table 2. Hausdorff Distance of Models on Different Organs

Organ \ Model	MSN	EMSN	EMSN*	EMSN*+
Esophagus	0.7095	0.3637	0.4416	**0.2883**
Heart	0.3733	0.2874	0.1696	**0.1594**
Trachea	0.5372	0.2670	0.2753	**0.2045**
Aorta	0.5221	0.2517	0.2040	**0.1551**
Mean	0.5355	0.2924	0.2727	**0.2018**

4. CONCLUSIONS

In this paper, we proposed a 3D 3-stage Enhanced Multi-scale Network (EMSN) to address the 4 organs at risk 3D data segmentation problem. Our network refines the prediction through a progressive auto-context procedure. The experiment results demonstrate that compared with baseline, the

overall performance on all classes improves a lot. Specially, the segmentation result of the hard class (esophagus) is also remarkably improved.

Acknowledgments

This work is supported by the Shenzhen Fundamental Research Fund under grants No. KQTD2015033114415450, No. ZDSYS201707251409055, and grant No. 2017ZT07X152.

5. REFERENCES

[1] National Lung Screening Trial Research Team, "Reduced lung-cancer mortality with low-dose computed tomographic screening," *New England Journal of Medicine*, vol. 365, no. 5, pp. 395–409, 2011.

[2] Ronneberger O, Fischer P, and Brox T, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[3] F. Milletari, N. Navab, and S. A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *Fourth International Conference on 3d Vision*. Springer, 2016.

[4] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger, "3d u-net: learning dense volumetric segmentation from sparse annotation," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2016, pp. 424–432.

[5] Eli Gibson, Francesco Giganti, Yipeng Hu, Ester Bonmati, Steve Bandula, Kurinchi Gurusamy, Brian Davidson, Stephen P. Pereira, Matthew J. Clarkson, , and Dean C. Barratt, "Automatic multi-organ segmentation on abdominal ct with dense v-networks," in *IEEE Transactions on Medical Imaging*. IEEE, 2018.

[6] Holger R.Roth, Chen Shen, Hirohisa Oda, Takaaki Sugino, Masahiro Oda, Yuichiro Hayashi, Kazunari Misawa, and Kensaku Mori, "A multi-scale pyramid of 3d fully convolutional networks for abdominal multi-organ segmentation," in *International Conference on Medical Image Computing and Computer Assisted Intervention*. Springer, 2018.

[7] Jonathan Long, Evan Shelhamer, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[8] Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *CoRR*, vol. abs/1207.0580, 2012.

[9] Fisher Yu and Vladlen Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[10] Bennett Landman et al, "Multi-atlas labeling beyond the cranial vault - workshop and challenge," https://www.synapse.org/#!Synapse:syn3193805/wiki/89480, February 14, 2015.

[11] The Centre for Medical Image Computing at University College London, "Niftyreg Software," http://cmictig.cs.ucl.ac.uk/wiki/index.php/NiftyReg, February 5, 2019.

[12] Daniel P Huttenlocher, William J Rucklidge, and Gregory A Klanderman, "Comparing images using the hausdorff distance under translation," in *Proceedings 1992 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 1992, pp. 654–656.