A 2D DILATED RESIDUAL U-NET FOR MULTI-ORGAN SEGMENTATION IN THORACIC CT

Sulaiman Vesal, Nishant Ravikumar, Andreas Maier

Pattern Recognition Lab, Friedrich-Alexander-University Erlangen-Nuremberg, Germany

ABSTRACT

Automatic segmentation of organs-at-risk (OAR) in computed tomography (CT) is an essential part of planning effective treatment strategies to combat lung and esophageal cancer. Accurate segmentation of organs surrounding tumours helps account for the variation in position and morphology inherent across patients, thereby facilitating adaptive and computer-assisted radiotherapy. Although manual delineation of OARs is still highly prevalent, it is prone to errors due to complex variations in the shape and position of organs across patients, and low soft tissue contrast between neighbouring organs in CT images. Recently, deep convolutional neural networks (CNNs) have gained tremendous traction and achieved state-of-the-art results in medical image segmentation. In this paper, we propose a deep learning framework to segment OARs in thoracic CT images, specifically for the: heart, esophagus, trachea and aorta. Our approach employs dilated convolutions and aggregated residual connections in the bottleneck of a U-Net styled network, which incorporates global context and dense information. Our method achieved an overall Dice score of 91.57% on 20 unseen test samples from the ISBI 2019 SegTHOR challenge.

***Index Terms*—** Thoracic Organs, Convolutional Neural Network, Dilated Convolutions, 2D Segmentation

1. INTRODUCTION

Organs at risk (OAR) refer to structures surrounding tumours, at risk of damage during radiotherapy treatment [1]. Accurate segmentation of OARs is crucial for efficient planning of radiation therapy, a fundamental part of treating different types of cancer. However, manual segmentation of OARs in computed tomography (CT) images for structural analysis, is very time-consuming, susceptible to manual errors, and is subject to inter-rater differences[1][2]. Soft tissue structures in CT images normally have very little contrast, particularly in the case of the esophagus. Consequently, an automatic approach to OAR segmentation is imperative for improved radiotherapy treatment planning, delivery and overall patient prognosis. Such a framework would also assist radiation oncologists in delineating OARs more accurately, consistently, and efficiently. Several studies have addressed automatic segmentation of OARs in CT images, with efforts being more focused on pelvic, head and neck areas [1][2][3].

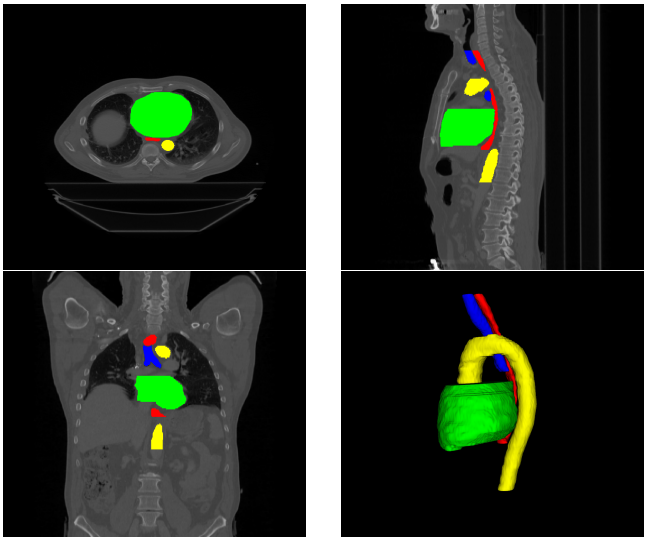

Fig. 1. Example of OARs in CT images with axial, sagittal and coronal views and 3D surface mesh plot.

In this paper, we propose a fully automatic 2D segmentation approach for the esophagus, heart, aorta, and trachea, in CT images of patients diagnosed with lung cancer. Accurate multi-organ segmentation requires incorporation of both local and global information. Consequently, we modified the original 2D U-Net [4], using dilated convolutions [5] in the lowest layer of the encoder-branch, to extract features spanning a wider spatial range. Additionally, we added residual connections between convolution layers in the encoder branch of the network, to better incorporate multi-scale image information and ensure a smoother flow of gradients in the backward pass.

2. METHODS

Segmentation tasks generally benefit from incorporating local and global contextual information. In a conventional U-Net [4] however, the lowest level of the network has a relatively

Thanks to EFI-Erlangen for funding.

small receptive field, which prevents the network from extracting features that capture non-local information. Hence, the network may lack the information necessary to recognize boundaries between adjacent organs, the fully connected nature of specific organs, among other properties that require greater global context to be included within the learning process. Dilated convolutions [5] provide a suitable solution to this problem. They introduce an additional parameter, i.e. the dilation rate, to convolution layers, which defines the spacing between weights in a kernel. This helps dilate the kernel such that a 3×3 kernel with a dilation rate of 2 results in a receptive field size equal to that of a 7×7 kernel. Additionally, this is achieved without any increase in complexity, as the number of parameters associated with the kernel remains the same.

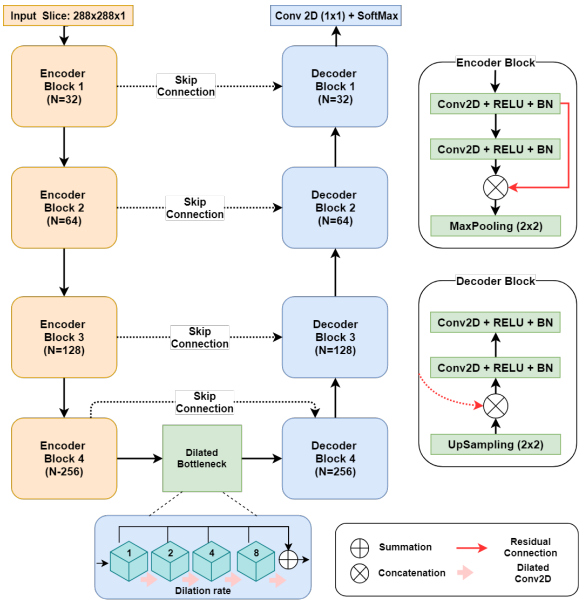

Fig. 2. Block diagram of the 2D U-Net+DR architecture for thoracic OAR images segmentation. The left side shows the encoding part and the right side shows the decoding part. The network has four dilated convolutions in the bottleneck and residual connection in each encoder block respectively(shown in red color arrow).

We propose a 2D U-Net+DR (refer to Fig.3.) network inspired by our previous studies [6][7]. It comprises four downsampling and upsampling convolution blocks within the encoder and decoder branches, respectively. In contrast to our previous approaches, here we employ a 2D version (rather than 3D) of the network with greater depth, because of the limited number of training samples. For each block, we use two convolutions with a kernel size of 3×3 pixels, with batch normalization, rectified linear units (ReLUs) as activation functions, and a subsequent max pooling operation. Image dimensions are preserved between the encoder-decoder branches following convolutions, by zero-padding the estimated feature maps. This enabled corresponding feature maps to be concatenated between the branches. A softmax activation function was used in the last layer to produce five probability maps to distinguish the background from the foreground labels. Furthermore, to improve the flow of gradients in the backward pass of the network, the convolution layers in the encoder branch were replaced with residual convolution layers. In each encoder-convolution block, the input to the first convolution layer is concatenated with the output of second convolution layer (red line in Fig. 3), and the subsequent 2D max-pooling layer reduces volume dimensions by half. The bottleneck between the branches employs four dilated convolutions, with dilation rates $1-4$. The outputs of each are summed up and provided as input to the decoder branch.

2.1. Dataset and Materials

The ISBI SegTHOR challenge[1] organizer provided the computed tomography (CT) images from the medical records of 60 patients. The CT scans are 512 × 512 pixels in size, with an in-plane resolution varying between 0.90 mm and 1.37 mm per pixel. The number of slices varies from 150 to 284 with a z-resolution between 2mm and 3.7mm. The most common resolution is 0.98×0.98×2.5 mm^3. The SegTHOR dataset (60 patients) was randomly split into a training set: 40 patients(7390 slices) and a testing set: 20 patients(3694 slices). The ground truth for OARs was delineated by an experienced radiation oncologist [2].

2.2. Pre-Processing

Due to low-contrast in most of CT volumes in the SegTHOR dataset, we enhanced the contrast slice-by-slice, using contrast limited adaptive histogram equalization (CLAHE), and normalized each volume with respect to mean and standard deviation. In order to retain just the region of interest (ROI), i.e. the body part and its anatomical structures, as the input to our network, each volume was center cropped to a size of 288×288 along the x and y axes, while the same number of slices along z were retained. We trained the model using the provided training samples via five-fold cross-validation (each fold comprising 32 subjects for training and 8 subjects for validation). Moreover, we applied off-line augmentation to increase the number of subjects within the training set, by flipping the volumes horizontally and vertically.

2.3. Loss Function

In order to train our model, we formulated a modified version of soft-Dice loss [8] for multiclass segmentation. Here the Dice loss for each class is first computed individually and then averaged over the number of classes. Let's suppose for the segmentation of an N×N input image (CT slice with esophagus, heart, aorta, trachea and background as labels), the out-

[1] https://competitions.codalab.org/competitions/21012

puts are five probabilities with classes of $k = 0, 1, 2, 3, 4$, such that $\sum_k \hat{y}_{n,k} = 1$ for each pixel. Correspondingly, if $y_{n,k}$ is the one-hot encoded ground truth of that pixel, then the multiclass soft Dice loss is defined as follows:

$$\zeta_{dc}(y, \hat{y}) = 1 - \frac{1}{N}(\sum_k \frac{\sum_n y_{nk}\hat{y}_{nk}}{\sum_n y_{nk} + \sum_n \hat{y}_{nk}}) \quad (1)$$

In Eq. (1) $\hat{y}_{nk}$ denotes the output of the model, where n represents the pixels and k denotes the classes. The ground truth labels are denoted by y_{nk}.

Furthermore, in the second stage of the training (described in detail in the next section), we used Tversky Loss (TL)[9], as the multiclass Dice loss does not incorporate a weighting mechanism for classes with fewer pixels. The TL is defined as following:

$$TL(y, \hat{y}) = 1 - \frac{\sum_{k=1}^{N} y_{nk}\hat{y}_{nk}}{\sum_{k=1}^{N} y_{nk}\hat{y}_{nk} + \alpha \sum_{k=1}^{N} y_{nk}\hat{y}_{nk} + \beta \sum_{k=1}^{N} y_{nk}\hat{y}_{nk}} \quad (2)$$

Also by adjusting the hyper-parameters α and β (refer to Eq. 2) we can control the trade-off between false positives and false negatives. In our experiments, we set both α and β to 0.5. Training with this loss for additional epochs improved the segmentation accuracy on the validation set as well as on the SegTHOR test set, compared to training with the multiclass Dice loss alone.

2.4. Model Training

The adaptive moment estimation (ADAM) optimizer was used to estimate network parameters throughout, and the 1st and 2nd-moment estimates were set to 0.9 and 0.999 respectively. The learning rate was initialized to 0.0001 with a decay factor of 0.2 during training. Validation accuracy was evaluated after each epoch during training, until it stopped increasing. Subsequently, the best performing model was selected for evaluation on the test set. We first trained our model using five-fold cross-validation without any online data augmentation and using only multiclass Dice loss function. In the second stage, in order to improve the segmentation accuracy, we loaded the weights from the first stage and trained the model with random online data augmentation (zooming, rotation, shifting, shearing, and cropping) for 50 additional epochs. This lead to significant performance improvement on the SegTHOR test data. As the multiclass Dice loss does not account for class imbalance, we further improved the second stage of the training process, by employing the TL in place of the former. Consequently, the highest accuracy achieved by our approach employed the TL along with online data augmentation. The network was implemented using Keras, an open-source deep learning library for Python, and was trained on an NVIDIA Titan X-Pascal GPU with 3840 CUDA cores and 12GB RAM. On the test dataset, we observed that our model predicted small structures in implausible locations. This was addressed by post-processing the segmentations, to retain only the largest connected component for each structure. As the segmentations predicted by our network were already of good quality, this only lead to marginal improvements in the average Dice score, of approximately 0.002. However, it substantially reduced the average Hausdorff distance, which is very sensitive to outliers.

2.5. Evaluation Metrics

Two standard evaluation metrics are used assess segmentation accuracy, namely, the Dice score coefficient (DSC) and Hausdorff distance (HD). The DSC metric, also known as F1-score, measures the similarity/overlap between manual and automatic segmentation. DSC metric is the most widely used metric to evaluate segmentation accuracy, and is defined as:

$$DSC(G, P) = \frac{2TP}{(FP + 2TP + FN)} = \frac{2|G_i \cap P_i|}{|G_i| + |P_i|} \quad (3)$$

The HD is defined as the largest of the pairwise distances from points in one set to their corresponding closest points in another set. This is expressed as:

$$HD(G, P) = \max_{g \in G} \left\{ \max_{p \in P} \left\{ \sqrt{g^2 - p^2} \right\} \right\} \quad (4)$$

In Eq. (3) and (4), (G) and (P) represent the ground truth and predicted segmentations, respectively.

3. RESULTS AND DISCUSSIONS

The average DSC and HD measures achieved by 2D U-Net+DR across five-fold cross-validation experiments are summarized in Table 1. The DSC scores achieved by the 2D U-Net+DR without data augmentation for the validation and test sets were 93.61% and 88.69%, respectively. The same network with online data augmentation significantly improved the segmentation accuracy to 94.53% and 91.43% for the validation and test sets, respectively. Finally, on fine-tuning the trained network using the TL we achieved DSC scores of 94.59% and 91.57%, respectively. Compared to [2], our method achieved DSC and HD scores of 85.67% and 0.30mm for the esophagus, the most difficult OAR to segment. Table 2. illustrates the DSC and HD scores of each individual organ for 2D U-Net+DR method with online augmentation and TL on test data set.

The images presented in Fig.3 help visually assess the segmentation quality of the proposed method on three test volumes. Here, the green color represents the heart, and the red, blue and yellow colors represent the esophagus, trachea,

Methods	Train Data	Validation Data	Test Data	
	DSC [%]	DSC [%]	DSC [%]	HD [mm]
2D U-Net + DR	0.9784	0.9361	0.8869	0.4461
2D U-Net + DR (Augmented)	0.9741	0.9453	0.9143	0.2536
2D U-Net + DR (Augmented) + TL	**0.9749**	**0.9459**	**0.9157**	**0.2500**

Table 1. The DSC and HD scores for training, validation and test dataset.

and aorta respectively. We obtained the highest average DSC value and HD for the heart and Aorta because of its high contrast, regular shape, and larger size compared to the other organs. As expected, the esophagus had the lowest average DSC and HD values due to its irregularity and low contrast, making it difficult to identify within CT volumes. However, our method achieved a DSC score of 85.8% for the esophagus on test data set, demonstrating better performance in comparison to the method proposed in [2] which used a shape mask network architecture and conditional random fields. These results highlight the effectiveness of the proposed approach for segmenting OARs, which is essential for radiation therapy planning.

Metrics	Esophagus	Heart	Trachea	Aorta
DSC [%]	0.858	0.941	0.926	0.938
HD [mm]	0.331	0.226	0.193	0.297

Table 2. The DSC and HD scores of each organ for 2D U-Net + DR(Augmented) + TL method.

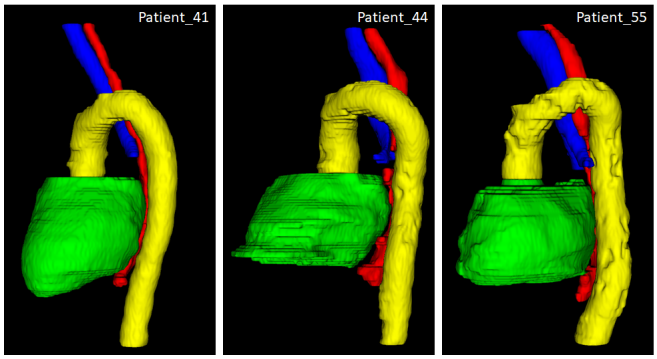

Fig. 3. 3D surface segmentation outputs of proposed model for three subjects from ISBI SegTHOR challenge test set.

4. CONCLUSIONS

In this study, we presented a fully automated approach, called 2D U-Net+DR, for automatic segmentation of the OARs (esophagus, heart, aorta, and trachea) in CT volumes. Our approach provides accurate and reproducible segmentations, thereby aiding in improving consistency and robustness in delineating OARs, relative to manual segmentations. The method uses both local and global information, by expanding the receptive-field in the lowest level of the network, using dilated convolutions. The two-stage training strategy employed here, together with the multi-class soft Dice loss and Tversky loss, significantly improved the segmentation accuracy. Furthermore, we believe that including additional information, e.g. MR images, may be beneficial for some OARs with poorly-visible boundaries such as the esophagus.

5. REFERENCES

[1] I. Bulat and L. Xing, "Segmentation of organs-at-risks in head and neck ct images using convolutional neural networks," *Medical Physics*, vol. 44, no. 2, pp. 547–557, 2017.

[2] R. Trullo, C. Petitjean, S. Ruan, B. Dubray, D. Nie, and D. Shen, "Segmentation of organs at risk in thoracic ct images using a sharpmask architecture and conditional random fields," in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, April 2017, pp. 1003–1006.

[3] S. Kazemifar, A. Balagopal, D. Nguyen, S. McGuire, R. Hannan, S. Jiang, and A. Owrangi, "Segmentation of the prostate and organs at risk in male pelvic CT images using deep learning," *Biomedical Physics & Engineering Express*, vol. 4, no. 5, pp. 055003, jul 2018.

[4] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, 2015, pp. 234–241.

[5] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *CoRR*, vol. abs/1511.07122, 2015.

[6] S. Vesal, N. Ravikumar, and A. Maier, "Dilated convolutions in neural networks for left atrial segmentation in 3d gadolinium enhanced-mri," in *STACOM. Atrial Segmentation and LV Quantification Challenges*, 2019, pp. 319–328.

[7] L. Folle, S. Vesal, N. Ravikumar, and A. Maier, "Dilated deeply supervised networks for hippocampus segmentation in mri," in *Bildverarbeitung für die Medizin 2019*, 2019, pp. 68–73.

[8] F. Milletari, N. Navab, and S. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 Fourth International Conference on 3D Vision (3DV)*, Oct 2016, pp. 565–571.

[9] S. Salehi, D. Erdogmus, and A. Gholipour, "Tversky loss function for image segmentation using 3d fully convolutional deep networks," in *Machine Learning in Medical Imaging*, 2017, pp. 379–387.