Rapid segmentation of thoracic organs using U-net architecture

Hassan Mahmood
School of Science
Edith Cowan University (ECU)
Joondalup, Australia
hmahmood@our.ecu.edu.au

Syed Mohammed Shamsul Islam
School of Science
Edith Cowan University (ECU)
Joondalup, Australia
syed.islam@ecu.edu.au

James Hill
Singular Health
Subiaco
WA, Australia
jhill@singular.health

Guan Tay
Singular Health
Subiaco
WA, Australia
gtay@singular.health

***Abstract*—Medical imaging provides a non-invasive method to diagnose, monitor and plan the treatment of disease inside the human body. The increasing prevalence of radiological scanners and prescription of their use has presented a significant challenge for radiologists in accurately diagnosing disease whilst dealing with a growing number of scans to review. Recent advances in Artificial Intelligence (AI), especially in machine learning, is enabling researchers to improve the patient experience, enhance the planning of medical treatments and increase the rate of examination of scans. In this study,a 2-dimensional (2D) U-net based deep learning model was used to automatically segment five organs of interest from Computed Tomography (CT) scans of the thoracic region. Comparable results were achieved in comparison to top seven models from a prior thoracic organ segmentation challenge. The framework can perform the segmentation tasks within 20 seconds, reducing workload for radiologists and increasing throughput. This study shows that a simple U-net framework can be sufficient for the task at hand rather than pursuing much more complicated architectures, depending upon the complexity of the problem. Furthermore, we investigated the effect of 3D interpolation on dice scores in anticipation of further research applications in mapping segments to a 3D volume render. We find performance degradation with respect to the dice score after mapping the masks to original dimensions.**

***Index Terms*—U-net, Thorax, Organs, Segmentation, Interpolation, Medical Imaging**

I. Introduction

One of the main challenges in the medical field is to diagnose disease and treat patients whilst minimising adverse impacts from the diagnostic/treatment procedure(s). For chronic and acute pathologies which present, and can impact, internal anatomy, medical imaging technologies allow clinicians to observe underlying phenomena inside the body without using invasive procedures. There are various imaging modalities for observing the different phenomenon in regions of the human body, e.g., Magnetic Resonance Imaging (MRI) [2], CT [5], Positron Emission Tomography (PET) [4], X-ray [3], ultrasound [6], etc. To facilitate diagnostic review of the imaging data, segmentation is one of the crucial tasks in medical imaging analysis. It provides the ability to segment the region(s) of interest (ROI) such as particular tissue or organs, benign or malignant tumors, etc. [1] and extract these ROIs for further review and even creating 3D printing biomodels for surgical planning. Despite being a crucial task, manual segmentation is tiresome and time-consuming, which makes accuracy susceptible to variation between different raters [7]. By incorporating AI and machine learning for the auto-segmentation algorithm, we can significantly reduce the computation time, increase accuracy, and provide increased capacity for the radiologist/rater. This work shows that a comparable performance can be achieved by standard U-net without complex architectural changes but by focusing on other parameters such as loss, data augmentation, number of layers, etc. The objective is to develop an automatic segmentation model for five Organs At Risk (OARs) in thoracic CT scans: heart, lungs(left and right), spinal cord, esophagus. Performance comparisons were made with the models reported in the AAPM 2017 challenge [8].

II. Related works

Earlier version of auto-segmentation algorithms were based on directly exploiting/modeling the morphological information in scans. These methods use intensity gradient and neighborhood structures around organs of interest to define the descriptors for boundaries [9]–[11], the performance of these methods is greatly dependant upon a consistent appearance. Besides methods based on mathematical modeling of the organs, atlas-based segmentation approaches are also common. These methods are relatively robust than the learning-based methods [12] however the performance of these methods is mainly subject to the quality of registration. Atlas-based

Identify applicable funding agency here. If none, delete this.

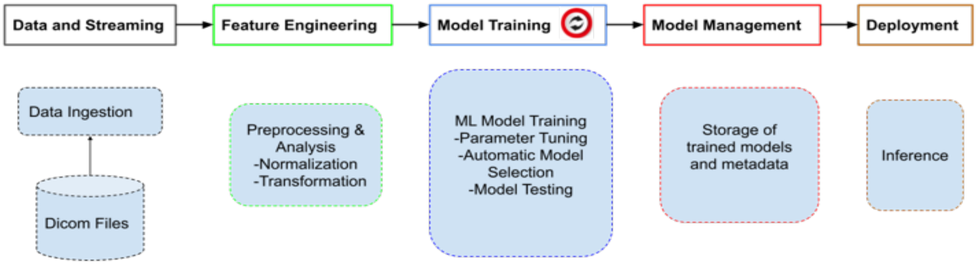

Data and Streaming
Feature Engineering
Model Training
Model Management
Deployment
Data Ingestion
Dicom Files
Preprocessing & Analysis -Normalization -Transformation
ML Model Training -Parameter Tuning -Automatic Model Selection -Model Testing
Storage of trained models and metadata
Inference

Fig. 1. Diagram of general machine learning pipeline from import of dataset through to deployment of trained model

segmentation methods take a high processing time due to recursive registration tasks. Deep learning (DL) based methods are making the registration process computationally efficient [13], [14]. Using these new models, atlas-based segmentation methods have become much faster however the quality of registration remains a core dependency for atlas-based segmentation, restricting its use.

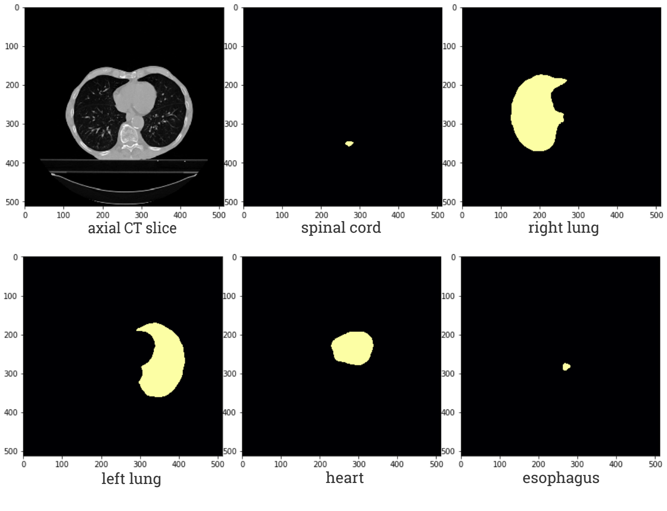
Among learning-based methods, DL-based segmentation methods are performing well. DL-based methods can learn features directly from large datasets. These methods are mostly based on convolutional neural networks (CNNs). Fully Convolutional Networks (FCNs) were first applied for image segmentation [15] but the results were relatively blurry, later the U-net [16] framework was proposed, which is a deep-learning model based on an encoder-decoder structure with the addition of skip connections. These skip connections are important to provide semantic information during training of the network. U-net got its popularity and many different alterations in architectures are proposed like 3D U-net [17], V-net [20], Dense U-net [18], MultiRes U-net [19], each showing better results compare to others. Due to differences in application, its difficult to make fair comparison [22]. [8] mentions the top seven best-performing methods used in the AAPM 2017 Challenge. This challenge's purpose was to evaluate the performance of segmentation methods for thoracic segmentation. Five of these methods are deep-learning based, while remaining two methods are multi-atlas based segmentation. The top-performing approach used a 2.5D model with an input size of 5x360x360 for lung segmentation and a 3D model with an input size of 32x128x128 was trained for the rest of the organs. The second-best performing method used a two-step strategy; first, a 3D U-net is used to locate the structures while the second 3D U-net is used to segment the organs. The third placed approach used a 2D multi-class network with fine-tuning of pre-trained network and loss function for small structures. The summary of these prior methods and their accuracy are provided in the results section and provide a contextual understanding of the results we achieved through our approach.

III. Method

A general ML pipeline is described in the Fig. 1. It shows different steps through which a final model gets ready to deploy. Some of the steps are covered in the following sections.

A. DataSet

A dataset from auto-segmentation for thoracic radiation treatment planning: a grand challenge at AAPM 2017 [8] is used. The purpose of this challenge was to compare different auto-segmentation models for five organs at risk (OARs) in the thoracic region. A total of 36 CT volumes are provided for training with a further 24 CT scans provided for testing and validation. This CT-Scan data is publicly available with segmentation masks of different thoracic regions including heart, lung (left), lung (right), spinal cord, esophagus. The scans in the dataset have varying slice thicknesses of 1 mm (MSKCC), 2.5 mm (MDACC), and 3 mm (MAASTRO). The number of slices is in the range from 103 to 279. Each slice is having a dimension of 512x512 pixels (px) . Representative sample images are shown in Fig. 2.



axial CT slice

spinal cord

right lung

left lung

heart

esophagus

Fig. 2. Sample axial slice and its respective masks for the five organs at risk.

B. Preprocessing and Data Augmentation

Preprocessing is one of the crucial parts of any machine learning setup and is used to standardise and improve data quality. During the preprocessing stage we cropped, normalised and finally resized the data to meet the model's input layer size. The data was normalised to use a specific Hounsfield Unity (HU) thresholding value and was resized from the original slice size of 512 x 512 (px) to 256 x 256 (px). Whilst the data has varying slice thicknesses, up to as much as 3mm, to reduce the interpolation error, we did not resize the data in the Z dimension. Whilst resizing in z-dimension may not have caused any degradation to the performance, we found that errors could be induced when registering the predicted masks back to the source data if slice thickness was normalised. The cause of these errors was not investigated as part of this research and is an avenue for further research. For training, we took 8 random axial slices for each batch, which, at an input size of 256px x 256px x 8 equates to 524,288 voxels. Random translation, rotation, and brightness variation transforms were applied as a data augmentation step.

C. Model

We used the famous U-net architecture. Originally developed for applications in microscopy, the U-net framework is a deep learning model based on an encoder-decoder structure with skip connections which is now regarded as a gold-standard for biomedical image segmentation. The encoder steps reduce the feature map in a step-wise manner to extract the abstract information from the input whereas, the decoder steps reconstruct the image guided by the loss function. In U-net architecture, there are skip connections, which are crucial in providing semantic information for the generation of the required output image during training and inference time. We have trained the network with the bottleneck of 1024 and 2048, and with cross-entropy and dice losses. We also observed the effect of reducing convolution layers in each encoding and decoding step. A generic structure of U-net is shown in Figure 3. The model was implemented through TensorFlow 2.2.0 package and trained on a GTX 1080 GPU [21]. Models are trained using categorical cross entropy and dice based loss function as shown in equation (1) and (2). Adam optimizer with learning rate of 0.0001 is used for 100 epochs.

$$\mathcal{L}_{ce}\left(\hat{y}, y\right) = -\sum_i y_i \log\left(\hat{y}_i\right) \tag{1}$$

$$\mathcal{L}_{dice}\left(\hat{y}, y\right) = 1 - \sum_i dice\left(\hat{y}_i, y_i\right) \tag{2}$$

whereas dice score is defined as in eq. (3)

$$dice\left(\hat{y}, y\right) = \frac{2 * |\hat{y} \cap y|}{|\hat{y}| + |y|} \tag{3}$$

IV. Results & Discussion

A. Quantitative & Qualitative Results

Quantitative results are shown in Table I. Dice, Hausdorff distance (HD) and mean surface distances (MSD) are used as metric. The table shows the resulting dice scores for different organs based on bottleneck size, convolutional layer, and losses. The first observation is that models do not perform well for organs with a small cross-sectional area or with less number of pixels when trained with cross-entropy loss, however models trained with dice-based loss perform better for small organs. This is due to the nature of the dice score which is tackling the issue of the number of pixels per class. Secondly, it can be observed that a reduction in the number of convolution layers in both encoding and decoding paths is detrimental to the performance. From Fig. 4, we can infer from the overall results that training with dice-based loss is better than cross-entropy, regardless of the small architectural changes. While the performance of the models with two convolution layers, trained with dice-based loss, are close to each other regardless of the size of the bottleneck, the best model in current experiments featured a bottleneck of 2048 with two convolution layers in encoding and decoding path trained with dice-based loss.

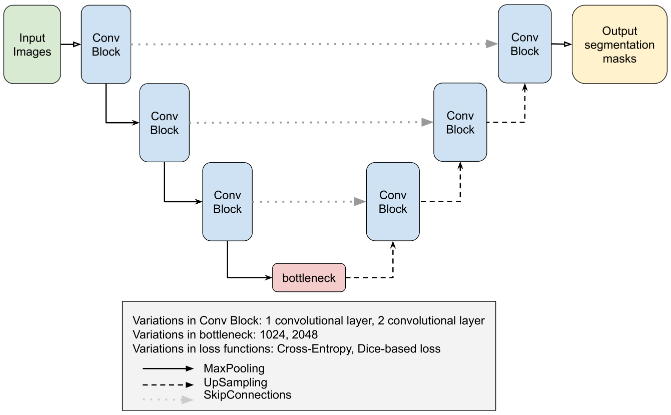

Input Images
Conv Block
Conv Block
Conv Block
bottleneck
Conv Block
Conv Block
Conv Block
Output segmentation masks
Variations in Conv Block: 1 convolutional layer, 2 convolutional layer
Variations in bottleneck: 1024, 2048
Variations in loss functions: Cross-Entropy, Dice-based loss
MaxPooling
UpSampling
SkipConnections

Fig. 3. Generic U-net architecture, with the variations mentioned in the legend. There are eight models trained and tested.

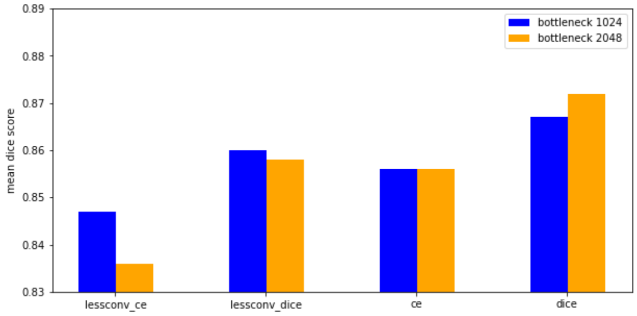


Fig. 4. mean dice score for each model based on ce and dice loss, number of convolution layers and bottleneck size

Table II, provides a comparative analysis of the different models from the AAPM 2017 Challenge. Despite the simple architecture, when comparing using Mean Dice score, our 2D U-net model, would have placed fourth in the Challenge.

TABLE I

SEGMENTATION PERFORMANCE OF THE EIGHT DIFFERENT VARIANTS OF U-NET FOR THE THORACIC ORGANS. THE RESULTS ARE EXPRESSED AS MEAN ± STANDARD DEVIATION FOR METRICS INCLUDING DICE COEFFICIENT, HAUSDORFF DISTANCE (HD), AND MEAN SURFACE DISTANCE (MSD)

Metric	**Model**	**Organ**				
		SpinalCord	Lung$_R$	Lung$_L$	Heart	Esophagus
Dice	1024 ce	0.79±0.07	0.97±0.01	0.97±0.01	0.91±0.03	0.63±0.10
	1024 dice	0.79±0.09	0.97±0.02	0.97±0.01	0.90±0.03	0.69±0.09
	1024 lessconv ce	0.76±0.11	0.96±0.02	0.96±0.01	0.90±0.03	0.63±0.08
	1024 lessconv dice	0.80±0.09	0.97±0.01	0.97±0.01	0.89±0.04	0.66±0.09
	2048 ce	0.79±0.09	0.96±0.02	0.97±0.01	0.91±0.02	0.63±0.11
	2048 dice	0.80±0.10	0.97±0.01	0.96±0.02	0.92±0.02	0.69±0.09
	2048 lessconv ce	0.77±0.10	0.97±0.01	0.97±0.01	0.89±0.05	0.56±0.15
	2048 lessconv dice	0.80±0.10	0.97±0.02	0.97±0.01	0.88±0.05	0.66±0.10
HD	1024 ce	19.4±17.0	22.6±29.1	22.3±35.3	9.90±10.5	9.41±5.19
	1024 dice	17.8±16.9	6.48±4.63	8.74±14.1	5.99±2.34	10.4±5.24
	1024 lessconv ce	37.8±28.9	49.8±23.5	44.4±35.1	10.1±9.64	20.1±17.7
	1024 lessconv dice	20.5±19.9	8.93±14.2	16.6±20.9	7.98±3.56	12.7±9.08
	2048 ce	17.0±15.3	9.79±7.89	7.78±6.79	8.26±4.66	11.4±7.63
	2048 dice	17.2±15.2	6.89±7.12	7.48±5.69	7.18±3.71	10.3±5.55
	2048 lessconv ce	17.6±15.0	5.87±3.58	8.35±12.1	6.91±3.73	10.2±5.59
	2048 lessconv dice	15.9±16.1	7.83±6.20	14.8±26.6	6.26±2.00	11.5±7.55
MSD	1024 ce	1.27±1.22	0.32±0.19	0.23±0.12	1.11±0.51	1.10±0.62
	1024 lessconv dice	1.55±1.73	0.32±0.22	0.28±0.19	1.04±0.46	0.89±0.57
	1024 lessconv ce	2.46±2.17	0.65±0.47	0.65±0.65	1.14±0.50	1.35±0.80
	1024 lessconv dice	1.47±1.72	0.28±0.19	0.30±0.27	1.26±0.57	1.05±0.70
	2048 ce	1.59±1.82	0.37±0.21	0.26±0.11	1.04±0.42	1.11±0.74
	2048 dice	1.65±1.89	0.29±0.19	0.31±0.28	0.92±0.35	0.94±0.63
	2048 lessconv ce	1.79±1.99	0.27±0.18	0.22±0.09	1.25±0.69	1.47±1.01
	2048 lessconv dice	1.56±1.79	0.31±0.20	0.29±0.24	1.30±0.62	1.09±1.03

TABLE II

COMPARISON AMONG TOP PERFORMING MODELS IN THE AAPM CHALLENGE

Method	Training Time	Inference Time	Run-time	Mean Dice
DL-1	3 days	30 sec	Titan X 12GB	0.889
DL-2	2 days	10 sec	Titan Xp 12GB	0.88
DL-3	>7 days	6 min	GTx 1050 2GB	0.88
Ours	**20 hrs**	**20 sec/ 35 sec**	**GTX 1080/ CPU 2.2 GHz**	0.87
MAC-1	-	8 hrs	-	0.87
DL-4	14 days	2 min	k40	0.85
MAC-2	-	5 min	-	0.85
DL-5	4 hrs	2 min	pascal	0.82

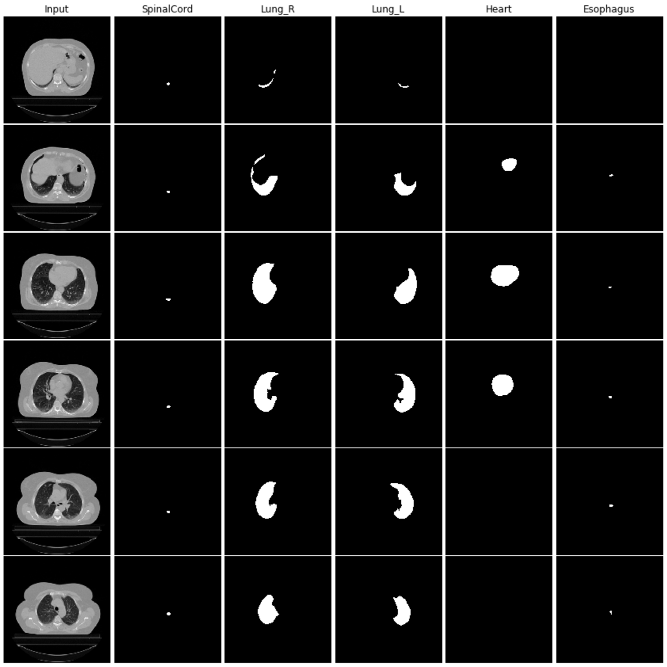


Fig. 5. Sample axial slices and its respective segmented masks for five organs.

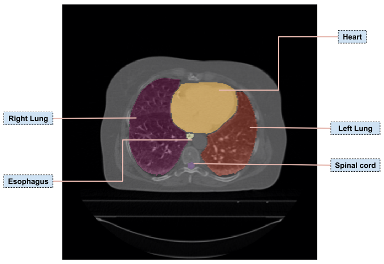


Fig. 6. Sample axial slice and its respective overlay masks for five organs.

Furthermore, we achieved this result with a training time that was less than one-third that of the top-performing methods and achieved an inference time which would have been second fastest on GPU-based runtime and third fastest on CPU-based runtime.

In Fig. 5, sample segmented masks for each OAR are shown in separate columns representing a series of individual CT images in the Z-plane. Fig. 6 shows the masks overlaid on a single gray scale input slice. Figure 7 demonstrates a 3D visualisation of the segmented organs.

B. Effects of Z-dimension Interpolation on Dice Score

We also explored the effect of interpolation on Dice scores, which involves generating interpolated images / slices to provide greater accuracy in the inferred masks. However, particularly for 3D U-net, input volumes need to be resized to match the fixed input size layer and to fit the model in GPU memory. Resizing the raw data in the Z-dimension is found to cause an issue, when attempting to register generated masks back to the raw data, we are not able to accurately generate the interpolated slices which are lost during the preprocessing stage. For the general task of population studies or diagnosis, using large slice thicknesses is recommended to minimise radiation exposure, however for surgery planning and radiotherapy, the source data must use smaller slice thicknesses so that the region of interest can be precisely segmented. Fig. 8 shows the trend, where the volume is first down-sampled to 132, 122, 112, 102, 92, 82, 72 slices in the Z-dimension and then resized to original depth. X and Y dimensions are kept the same at 512 x 512 pixels. The Dice score is computed between original volume and resized volume and it is possible to see a clear decrease in dice score as we move towards the right in the bar graph. The table in figure 8 shows the average dice score of each organ, whereas the bar plot shows mean dice score over all five organs for each case separately.

V. CONCLUSION & FUTURE DIRECTIONS

In conclusion, we show that a simple deployment of 2D U-net can deliver comparable performance with a less complex architecture for thoracic organ segmentation which can be deployed on a single GPU or CPU. The combination of relatively light and fast model is good for real-time patient handling, speed-up the radiologist task without demand of high computational resources. Overall 2D U-net based framework is

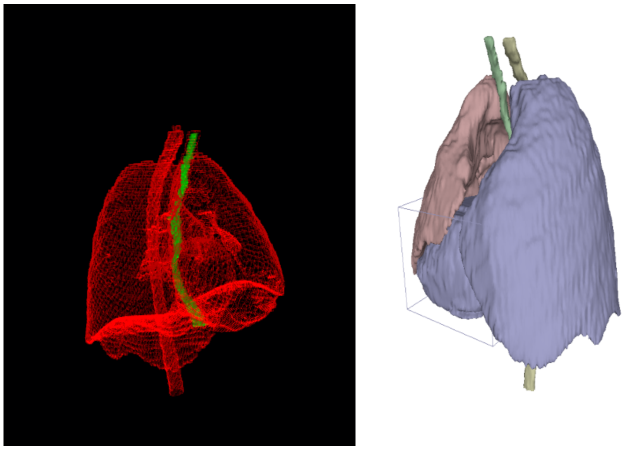
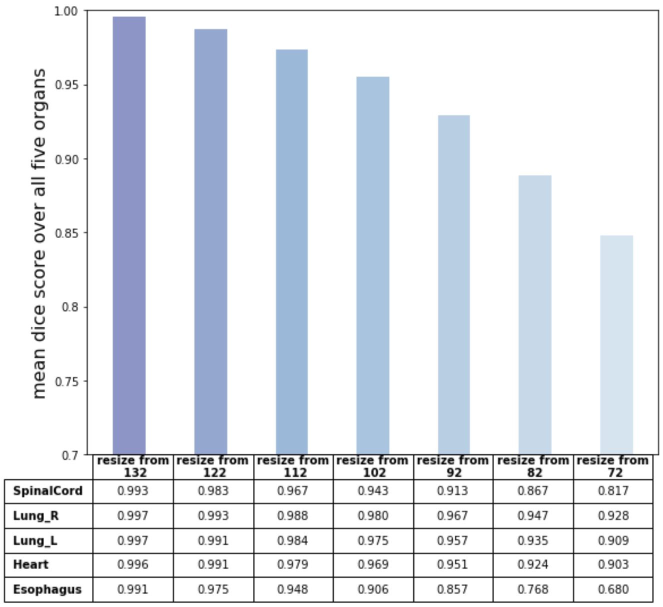


Fig. 7. (left) Raw 3D plot of segmented organs, (right) 3D STL file view of segmented organs



	resize from 132	resize from 122	resize from 112	resize from 102	resize from 92	resize from 82	resize from 72
SpinalCord	0.993	0.983	0.967	0.943	0.913	0.867	0.817
Lung_R	0.997	0.993	0.988	0.980	0.967	0.947	0.928
Lung_L	0.997	0.991	0.984	0.975	0.957	0.935	0.909
Heart	0.996	0.991	0.979	0.969	0.951	0.924	0.903
Esophagus	0.991	0.975	0.948	0.906	0.857	0.768	0.680

Fig. 8. Effect of interpolation in Z-dimension on dice score

capable of attaining relative results in comparison with the top seven methods rated by the AAPM Challenge. We demonstrate that comparative performance can be achieved without major architectural changes but with the right choice of the loss function and the addition or removal of convolutional layers. Complex architectural changes can be made, and they can be effective depending upon the complexity of the task. In [8] and in our experimentation , dice score range for esophagus is between 0.55-0.72, while the inter-rater difference in dice score for esophagus is 0.81. The exact reason for the reduced dice score across all the models need further investigation. It can be due to the class pixel imbalance or due to less contrast with the surrounding tissues as compare to the spinal cord. Although both have almost same cross sectional area, spinal cord is having relatively much better dice score, due high contrast with its surroundings.

We further explore the effect of resizing/interpolation of 3D CT-scan on dice score. For diagnosis and population studies, it may not cause any problem. But, for precise tasks such as radiotherapy, removal of a tumor, surgery planning, etc., it would be favorable to avoid these errors and precise masks which would accurately overlay the original image are desirable [23]. We also notice a degradation in performance if resize the slice itself in the x-y plane, without any resizing in z-dimension. We propose that high-resolution segmentation frameworks or frameworks which are independent of resizing the volumes would be a possible future research direction, targeting towards the improvement of treatment procedures.

Acknowledgment

Authors would like to acknowledge that this research was conducted with the support of APR.Intern, Singular Health Group Limited, Edith Cowan University's School of Science 2017 Strategic Initiative Fund and Higher Education Commission of Pakistan (HEC)-ECU joint scholarship.

References

[1] Lei, Tao, et al. ”Medical image segmentation using deep learning: a survey.” arXiv preprint arXiv:2009.13120 (2020).

[2] Schaefer, Pamela W., P. Ellen Grant, and R. Gilberto Gonzalez. ”Diffusion-weighted MR imaging of the brain.” Radiology 217.2 (2000): 331-345.

[3] Ambrose, James. ”Computerized x-ray scanning of the brain.” Journal of neurosurgery 40.6 (1974): 679-695.

[4] Svarer, Claus, et al. ”MR-based automatic delineation of volumes of interest in human brain PET images using probability maps.” Neuroimage 24.4 (2005): 969-979.

[5] Buzug, Thorsten M. ”Computed tomography.” Springer handbook of medical technology. Springer, Berlin, Heidelberg, 2011. 311-342.

[6] Fenster, Aaron, Donal B. Downey, and H. Neale Cardinal. ”Three-dimensional ultrasound imaging.” Physics in medicine biology 46.5 (2001): R67.

[7] Tingelhoff, Kathrin, et al. ”Analysis of manual segmentation in paranasal CT images.” European archives of oto-rhino-laryngology 265.9 (2008): 1061-1070.

[8] Yang, Jinzhong, et al. ”Autosegmentation for thoracic radiation treatment planning: a grand challenge at AAPM 2017.” Medical physics 45.10 (2018): 4568-4581.

[9] Kohlberger, Timo, et al. ”Automatic multi-organ segmentation using learning-based segmentation and level set optimization.” International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Berlin, Heidelberg, 2011.

[10] Kang, Dong Joong. ”A fast and stable snake algorithm for medical images.” Pattern Recognition Letters 20.5 (1999): 507-512.

[11] Duquette, Anthony Adam, et al. ”3D segmentation of abdominal aorta from CT-scan and MR images.” Computerized Medical Imaging and Graphics 36.4 (2012): 294-303.

[12] Zhuang, Xiahai, and Juan Shen. ”Multi-scale patch and multi-modality atlases for whole heart segmentation of MRI.” Medical image analysis 31 (2016): 77-87.

[13] Mahmood, Hassan, Asim Iqbal, and Syed Mohammed Shamsul Islam. ”Exploring Intensity Invariance in Deep Neural Networks for Brain Image Registration.” 2020 Digital Image Computing: Techniques and Applications (DICTA). IEEE, 2020.

[14] Balakrishnan, Guha, et al. ”Voxelmorph: a learning framework for deformable medical image registration.” IEEE transactions on medical imaging 38.8 (2019): 1788-1800.

[15] Long, Jonathan, Evan Shelhamer, and Trevor Darrell. ”Fully convolutional networks for semantic segmentation.” Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[16] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. ”U-net: Convolutional networks for biomedical image segmentation.” International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, 2015.

[17] Çiçek, Özgün, et al. ”3D U-Net: learning dense volumetric segmentation from sparse annotation.” International conference on medical image computing and computer-assisted intervention. Springer, Cham, 2016.

[18] Cai, Sijing, et al. ”Dense-UNet: a novel multiphoton in vivo cellular image segmentation model based on a convolutional neural network.” Quantitative imaging in medicine and surgery 10.6 (2020): 1275.

[19] Ibtehaz, Nabil, and M. Sohel Rahman. ”MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation.” Neural Networks 121 (2020): 74-87.

[20] Milletari, Fausto, Nassir Navab, and Seyed-Ahmad Ahmadi. ”V-net: Fully convolutional neural networks for volumetric medical image segmentation.” 2016 fourth international conference on 3D vision (3DV). IEEE, 2016.

[21] ”Deep Learning Workstations, Servers, Laptops For 2021 — Lambda”. Lambdalabs.Com, 2021, https://lambdalabs.com/products/blade.

[22] Feng, Xue, et al. ”Deep convolutional neural network for segmentation of thoracic organs-at-risk using cropped 3D images.” Medical physics 46.5 (2019): 2169-2180.

[23] Martin, Spencer, et al. ”Impact of target volume segmentation accuracy and variability on treatment planning for 4D-CT-based non-small cell lung cancer radiotherapy.” Acta Oncologica 54.3 (2015): 322-332.