MPHY0041 Coursework 2

Group E

Student No: 18035489, 18102269, 18044982, 22235817, 18012615

1 Introduction

Prostate cancer is one of the most prevalent cancers among men worldwide, with incidence and mortality rates being positively correlated with age [1]. Over the years, prostate cancer treatment has exhibited improved outcomes following prostatectomy and radiation therapy, as evidenced by promising overall 5-year survival rates ranging from 83% in Europe [2] to 98% in the USA [3].

In patients with high-risk prostate cancer, studies have shown that using multi-parametric MRI (mpMRI) can have a large positive impact on surgical planning [4,5]. mpMRI allows for accurate segmentation and characterisation of the tumour site, along with the detection of extraprostatic extension, including seminal vesicle and neurovascular bundle invasion [4]. This can lead to better decision-making for nerve-sparing procedures like bladder neck dissection, and ultimately lead to improved tumour management and quality of life [4].

The automatic feature representation learning enabled by deep learning (DL) neural networks has led to a revolutionary shift in the field of image segmentation. Numerous DL architectures have achieved remarkable performance and accuracy [6]. Automated segmentation offers advantages such as speed, robustness, and reduced labour compared to manual segmentation. Nevertheless, most segmentation networks are based on supervised methods and so rely on large, labelled datasets, whose availability often poses a challenge, especially when dealing with medical images.

One successful architecture that has been used extensively for medical image segmentation is the U-Net [7]. It has been shown to achieve good segmentation performance even when labelled training data is limited [8]. The negative impact of data sparsity can also be minimised by using multi-task learning (MTL). This learning paradigm assumes that generalisation performance on the primary task can be enhanced by simultaneously learning multiple other related tasks [9]. By using extensive data from diverse auxiliary tasks more powerful models can be created, therefore increasing the accuracy of the results [9].

According to research, multi-task networks that perform simultaneous automatic organ segmentation and image transfer have proved to be highly efficient and useful for enhancing the quality of both tasks. Playout et. al proposed an extended U-Net architecture that utilises multi-task learning for joint segmentation of bright and red lesions in fundus images [10]. Their approach involved a single encoding module and several decoding branches specialised for each category, which were able to capture features that a single branch could not. Wang et. al implemented a multi-task fully convolutional network (FCN) for concurrent segmentation of the lungs, heart, and clavicles in chest radiographs, achieving better results than using multiple FCNs or a single FCN with a combined objective function [11]. Sui et. al investigated the use of image reconstruction

by connecting two U-Nets for simultaneous MRI reconstruction, and liver and renal lesion segmentation [12]. Reconstructing the input image as one of the auxiliary tasks helped regularise the shared encoder, and hence improved the generalisation ability of the model [13]. Similarly, Weninger et. al highlighted the value of multi-task learning in 3D MRI brain tumour segmentation, by incorporating image reconstruction and detection of enhancing tumour tissue using a shared encoder [14].

In this study, we explored multi-task learning for planning prostate cancer interventions on pelvic T2-weighted (T2-w) MR images. The main task involved segmenting the central gland (CG) and the transition zone (TZ) of the prostate. Accurate segmentation of the prostate is critical for guiding surgical plans, as underestimation of the tumour burden can lead to incomplete resection and suboptimal treatment outcomes. It is also important to mitigate damage to the surrounding anatomy [15]. We tested two different auxiliary tasks and evaluated the ability of each to assist the main segmentation task. The first consisted of segmenting additional regions of interest (ROIs) (we tested three and six additional organs). This has clinical relevance because avoiding vulnerable organs surrounding the prostate is important for minimising the risk of impotence, infection, and other complications during surgery and radiotherapy. The second auxiliary task tested was image reconstruction. This was chosen with the aim of making the network more generalisable and robust to inter-scanner variability.

2 Methods

In this study, we utilised a publicly available dataset from UCL consisting of 589 T2-weighted MR images of the male pelvic region with eight anatomical structures segmented [15]. Since the images were acquired across different institutions using different imaging protocols, all images and their corresponding labels were resized to a common size 256 x 256 x 40. The intensity of all the training images was also rescaled to the range [0, 1].

Our base model consisted of an optimised 3D U-Net [16] trained to segment the prostate CG and TZ. The depth of our network was five, with convolution channels at each encoder level of [32, 64, 128, 256, 512]. The optimisations performed on the U-Net were inspired by the variety of network designs reviewed by Futrega et al. [17]. Firstly, we added residual connections within each convolution block; residual connections allow for deeper penetration of gradients during back-propagation and thus improve the efficiency of network training. Secondly, we used a Swish activation function within each convolution block. The primary advantage of the Swish activation function over ReLU is that the Swish function is differentiable at all points which helps to reduce the likelihood of vanishing and exploding gradients [18]. The convolution blocks also included a dropout layer [19] to prevent complex co-adaptations on training data hence reducing overfitting. Finally, we added a gated-attention mechanism in each skip connection which has been shown to allow U-Nets to focus more on certain

regions of the input image that are more relevant to the segmentation task [20]. Figure 1 displays a diagram of our model architecture.

We adapted our base 3D U-Net to simultaneously perform auxiliary tasks by adding additional decoder branches. Each decoder branch had an associated loss function relevant to its task. The task for the main decoder in all the trained models was the segmentation of the TZ and CG of the prostate. During training, the total network loss was set as a weighted sum of losses from each decoder branch with weightings of 1.5 and 1.1 for the main and auxiliary losses respectively. The first auxiliary task we implemented was the segmentation of additional ROIs surrounding the prostate. We separately trained two MTL models to segment three and six additional ROIs, which we refer to as *MTL Seg 3* and *MTL Seg 6* respectively. The loss function for the auxiliary decoder branch in these models was set as the Dice dissimilarity index between the input and output segmentation masks. The second auxiliary task we implemented was image reconstruction (*MTL Recon*), where the associated decoder branch loss function was set as the mean absolute error between the input and reconstructed images.

Due to memory limitations, we used a batch size of two during training. As a result, group normalisation was used instead of batch normalisation to speed up and stabilize training because batch normalisation has been shown to perform poorly with small batch sizes [21,22]. To increase robustness to noise and contrast variability, we artificially extended our training set by randomly adding rotations, scaling, Gaussian noise, Gaussian smoothing and altering the contrast of training images.

Our proposed models were written in PyTorch [23] and are publicly available on our GitHub repository. We make our trained model weights available for download here.

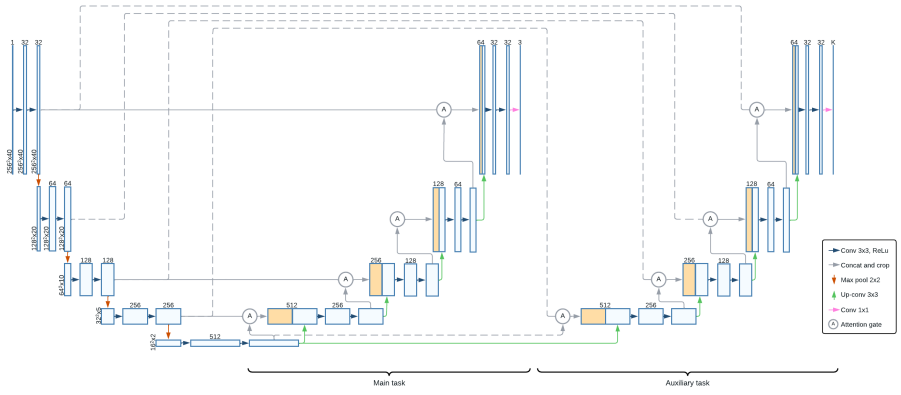

Fig. 1: Proposed 3D U-Net architecture

3 Experiments

To test the hypothesis of whether MTL improves the segmentation of the main prostate zones, four models were trained: a 3D-U-Net with one decoder which segments the TZ and CG and three 3D-U-Nets with two decoders (*MTL Seg 3*, *MTL Seg 6*, *MTL Recon*). We used a train-validation-test split percentage of 80%/10%/10%. We trained each model with 100 epochs using an Nvidia RTX A6000 GPU and it took approximately five hours to train each model. A learning rate of 0.001 was used and was progressively decreased using a cosine annealing scheduler and Adam as the optimiser [24]. After training, we evaluated the main task segmentation performance for each model using the Dice similarity coefficient between the output segmentation masks and the segmentation masks accompanying the test set; the Dice coefficient is a widely used metric for segmentation performance evaluation and is consistent with the loss functions we used during training. To evaluate our results, we performed a paired student's t-test between the distribution of results from the base model and each MTL model to identify any significant changes to the main task's performance.

4 Results

The Dice scores achieved for each of the variations in model architecture in this study are presented. Figure 2 displays the distribution of these similarity coefficients achieved on each individual in the testing set, including the performance of the main task alone as a base comparison. Here, the boxplot whiskers indicate the 5th and 95th percentiles of the performance score across the test set for each model. Overall, the MTL Seg 6 model outperformed the rest in segmenting both of the main ROIs, with a mean Dice score of 0.816 (0.867 for CG and 0.764 for TZ), whereas the base model achieved 0.803 (0.850 for CG and 0.756 for TZ). Notably, this was a statistically significant increase in performance for both CG and TZ segmentation with t-test p values of 0.00494 and 0.00178 respectively. No significant change to CG or TZ segmentation was present across the other model architectures. Overall, CG segmentation had a higher performance irrespective of the model used. Further, there was less variability in the Dice coefficients relative to the TZ, implied by the consistently smaller interquartile ranges. Figure 3 depicts the best and worst predictions from the test set across each of the different models and includes the Dice score found for each zone. It is evident that the models performed well compared to the ground truth when there was good contrast between the intensities of each section however, in some instances, all of the models had large errors when segmenting the TZ.

5 Discussion and Conclusion

We have explored the effect of exploiting different auxiliary tasks in MTL to segment key prostate regions using the Cross-institution Male Pelvic Structures dataset [15]. The best-performing model, MTL Seg 6, achieved the highest Dice

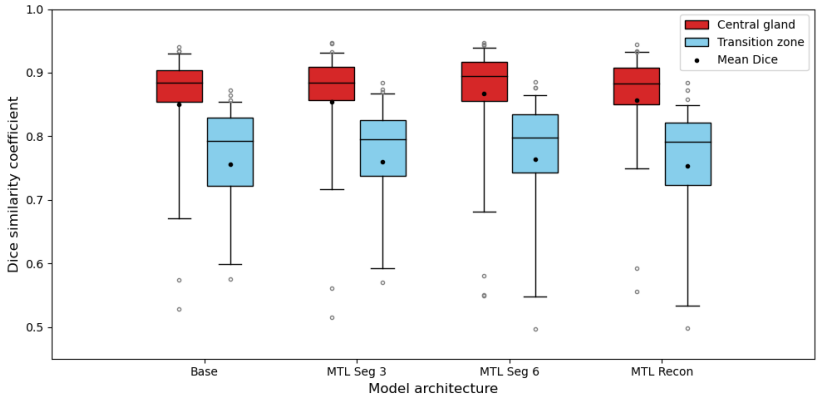

Fig. 2: Dice similarity coefficients for main task segmentation with different MTL U-Net architectures.

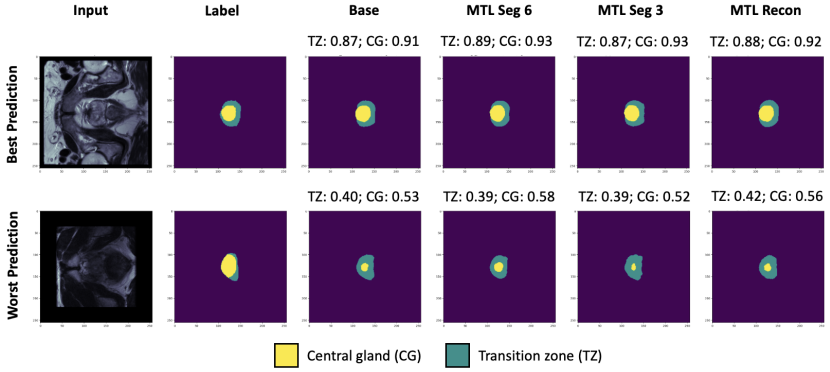

Fig. 3: Best and worst case predictions on the test set for each model, including the Dice score for each zone (CG and TZ). Darker images result in poorer differentiation.

score. Here, it is apparent that the information provided by the surrounding anatomical regions is useful for improving the accuracy of the main task. It was found that segmenting more anatomical regions (i.e., six instead of three) in the auxiliary task led to an improvement in performance, which can be attributed to numerous advantages. Firstly, the inclusion of more regions leads to an increase in contextual information, allowing the model to produce more precise segmentations. By considering more structures, the model can also extract more generalised features that can be shared across tasks, which can further mitigate overfitting the main task. Finally, this heightens robustness because the model can become acclimatised to variations in the data, especially from

different institutions. Nonetheless, the computational burden of including more segmentations should be considered as a trade-off with performance.

We found that the inclusion of image reconstruction as a supplementary task does not improve the performance in segmenting the CG and TZ. This implies that reconstruction does not provide as much useful information as additional segmentations; the task may have introduced additional noise and inaccuracies that could have impacted the main task's quality.

Regarding class performance, the models repeatedly segmented the CG with higher accuracy. This is consistent across other studies [25,26], whereby this gland is more successfully segmented in other U-Net-like architectures. This could be attributed to the clearer anatomical distinction of the CG, where its boundaries are more apparent in the scans, and it also appears to be more homogeneous across slices and patients. This is further reflected in Figure 2, where the CG has less disparity. Moreover, the difference in T2-w contrast between the two structures potentially leads to easier differentiation of the CG (it has higher intensities). On the whole, this is a positive outcome because T2-w sequences are the most important in determining cancer presence in the CG [25].

Our results are comparable with those found in other studies. Zabihollahy et. al [25] achieved a maximum Dice score of 0.97 when using a U-Net to find the prostate boundaries from T2-w images. However, this was from a single-institution cohort. Conversely, Tian et al. [27] trained a CNN to achieve a Dice score of 0.85 for three independent datasets, which is more analogous to our cross-institution dataset. Further, we achieved very similar performance to prior work on the same dataset [15], whereby few-shot segmentation was used to achieve the best scenario Dice score of 0.831.

Despite achieving reasonable results, there are some limitations of this work that should be explored in future studies. First, the training time should be increased to include more epochs. Further, test-time augmentation could be implemented to improve segmentation accuracy, as seen in prior work with MRI segmentation [28], and to enhance robustness. Moreover, hyperparameter tuning should be explored to ensure the correct weighting between the main and auxiliary tasks during learning, as well as exploring different loss functions (e.g. focal loss [29]). Different architectural changes could also be implemented, including deeper architecture [30] and deep supervision [17].

In summary, this work uses supervised multi-task learning with the main aim of segmenting the CG and TZ of the prostate to assist in surgical planning. Multiple auxiliary tasks were considered, whereby it was found that segmenting six other pelvic anatomical regions was the most useful for improving performance. The model was trained on pelvic T2-w MRI images from different institutes. The demonstrated high performance can therefore be linked to the generalisability of the model. Finally, it is notable that the automatic segmentation pipeline presented could help alleviate the burden of manual segmentation; the dataset itself required an excess of one-thousand observer-hours to produce [15], emphasising the need for this work.

References

1. Prashanth Rawla. Epidemiology of Prostate Cancer. 10(2):63–89.
2. Maximilian Marhold, Gero Kramer, Michael Krainer, and Clémentine Le Magnen. The prostate cancer landscape in Europe: Current challenges, future opportunities. 526:304–310.
3. A.M. Noone, N. Howlader, and M. Krapcho. SEER Cancer Statistics Review. `https://seer.cancer.gov/csr/1975_2015/`. Online; Accessed: 2023-04-25.
4. Janet Baack Kukreja, Tharakeswara K. Bathala, Chad A. Reichard, Patricia Troncoso, and Scott Delacroix. Impact of preoperative prostate magnetic resonance imaging on the surgical management of high-risk prostate cancer. 23(1):172–178.
5. J. De La Rosette, H. Ahmed, J. Barentsz, T. Bjerklund Johansen, M. Brausi, and M. Emberton. Focal Therapy in Prostate Cancer—Report from a Consensus Panel. 24(5):775–780.
6. Shervin Minaee, Yuri Boykov, Fatih Porikli, Antonio Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey.
7. Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.
8. Reza Azad, Ehsan Khodapanah Aghdam, Amelie Rauland, Yiwei Jia, Atlas Haddadi Avval, Afshin Bozorgpour, Sanaz Karimijafarbigloo, Joseph Paul Cohen, Ehsan Adeli, and Dorit Merhof. Medical image segmentation review: The success of u-net.
9. Yu Zhang and Qiang Yang. A survey on multi-task learning.
10. Clément Playout, Renaud Duval, and Farida Cheriet. A multitask learning architecture for simultaneous segmentation of bright and red lesions in fundus images. In Alejandro F. Frangi, Julia A. Schnabel, Christos Davatzikos, Carlos Alberola-López, and Gabor Fichtinger, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pages 101–108, Cham, 2018. Springer International Publishing.
11. Chunliang Wang. Segmentation of multiple structures in chest radiographs using multi-task fully convolutional networks. In Puneet Sharma and Filippo Maria Bianchi, editors, *Image Analysis*, pages 282–289, Cham, 2017. Springer International Publishing.
12. Bin Sui, Jun Lv, Xiangrong Tong, Yan Li, and Chengyan Wang. Simultaneous image reconstruction and lesion segmentation in accelerated MRI using multitasking learning. 48(11):7189–7198.
13. Andriy Myronenko. 3d MRI brain tumor segmentation using autoencoder regularization.
14. Leon Weninger, Qianyu Liu, and Dorit Merhof. Multi-task learning for brain tumor segmentation. In Alessandro Crimi and Spyridon Bakas, editors, *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, volume 11992, pages 327–337. Springer International Publishing. Series Title: Lecture Notes in Computer Science.
15. Yiwen Li, Yunguan Fu, Iani Gayo, Qianye Yang, Zhe Min, Shaheer Saeed, Wen Yan, Yipei Wang, J. Alison Noble, Mark Emberton, Matthew J. Clarkson, Henkjan Huisman, Dean Barratt, Victor Adrian Prisacariu, and Yipeng Hu. Prototypical few-shot segmentation for cross-institution male pelvic structures with spatial registration, 2022.

16. Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*, pages 424–432. Springer, 2016.
17. Michał Futrega, Alexandre Milesi, Michał Marcinkiewicz, and Pablo Ribalta. Optimized u-net for brain tumor segmentation. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 7th International Workshop, BrainLes 2021, Held in Conjunction with MICCAI 2021, Virtual Event, September 27, 2021, Revised Selected Papers, Part II*, pages 15–29. Springer, 2022.
18. Prajit Ramachandran, Barret Zoph, and Quoc V Le. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017.
19. Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.
20. Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.
21. Saurabh Singh and Shankar Krishnan. Filter response normalization layer: Eliminating batch dependence in the training of deep neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11237–11246, 2020.
22. Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.
23. Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.
24. Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
25. Fatemeh Zabihollahy, Nicola Schieda, Satheesh Krishna Jeyaraj, and Eranga Ukwatta. Automated segmentation of prostate zonal anatomy on t2-weighted (t2w) and apparent diffusion coefficient (scpADC/scp) map scpMR/scp images using u-nets. *Medical Physics*, 46(7):3078–3090, May 2019.
26. Nader Aldoj, Federico Biavati, Florian Michallek, Sebastian Stober, and Marc Dewey. Automatic prostate and prostate zones segmentation of magnetic resonance images using DenseNet-like u-net. *Scientific Reports*, 10(1), August 2020.
27. Zhiqiang Tian, Lizhi Liu, Zhenfeng Zhang, and Baowei Fei. PSNet: prostate segmentation on MRI based on a convolutional neural network. *Journal of Medical Imaging*, 5(02):1, January 2018.
28. Guotai Wang, Wenqi Li, Michael Aertsen, Jan Deprest, Sebastien Ourselin, and Tom Vercauteren. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. 2018.
29. Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection, 2017.
30. Qikui Zhu, Bo Du, Baris Turkbey, Peter L . Choyke, and Pingkun Yan. Deeply-supervised cnn for prostate segmentation, 2017.