Improving Prostate Segmentation through Hierarchical Loss Penalisation

MPHY0041 CW2 Group F: Charvi Agarwal, Ben Wright, Ishan Vermani, Yuanzhe Chang, Christos Toilos, Alexandros Mathios

University College London

We acknowledge the use of Anthropic Claude and OpenAI ChatGPT for brainstorming, code troubleshooting, and report proofreading.

1 Introduction

Machine learning techniques are widely used in medical image analysis, particularly for segmentation in imaging modalities such as CT and MRI [1]. In clinical applications such as prostate cancer radiotherapy, accurate segmentation of the prostate and surrounding pelvic organs from MRI scans is essential for treatment planning, as these delineations define target volumes and organs at risk. However, manual contouring is time-consuming and subject to inter-observer variability, motivating the use of automatic segmentation methods based on deep learning [5, 6].

Existing segmentation models are primarily trained using flat loss functions, such as cross-entropy or Dice-based losses, which assume that all misclassifications are equal. While this simplification facilitates optimisation, it does not reflect clinical reality, where different segmentation errors can have varying consequences. In clinical segmentation tasks, minor confusions between anatomically related structures may be acceptable, whereas anatomically invalid errors can compromise downstream clinical decision-making.

Loss penalisation strategies weigh segmentation errors according to their relative severity, capturing these differences. Hierarchical loss functions achieve this by encoding structured relationships between classes, enabling models to distinguish varying degrees of misclassification. Prior work has demonstrated the effectiveness of cost-sensitive loss formulations in related settings, including fine-grained segmentation and disease grading tasks [4, 2].

However, the potential of hierarchical loss remains underexplored in anatomically complex medical segmentation problems involving multiple organs. In the male pelvic region, MRI scans exhibit high anatomical variability, overlapping tissue boundaries, and subtle intensity differences between neighbouring structures. Moreover, pelvic organs play distinct roles in radiotherapy planning, and segmentation errors affecting critical structures such as the prostate can directly influence dose delivery and treatment safety. In such applications, treating all segmentation errors uniformly may obscure clinically important distinctions.

This work investigates whether incorporating a hierarchical loss into a pelvic organ segmentation framework improves validation performance compared to a

flat loss baseline. We evaluate this approach on prostate segmentation in pelvic MRI images by directly comparing hierarchical and flat loss strategies under otherwise identical experimental conditions, assessing whether anatomically informed loss design leads to measurable performance gains.

2 Methods

For multi-class segmentation, a 2D U-Net architecture was implemented because of its ability to accurately perform pixel-wise classification of complex anatomy in biomedical images [7]. The network consists of three encoding stages, where the number of feature channels increases from 64 to 256. In the decoding path, feature maps are upsampled using bilinear interpolation and combined with corresponding outputs through skip connections, enabling the recovery of fine spatial details. Each stage consists of two convolutional layers with 3×3 kernels and ReLU activations. A final 1×1 convolutional layer maps the extracted features to the output classes for pixel-wise prediction. U-Net was utilised to provide a strong baseline for both the flat and hierarchical loss models.

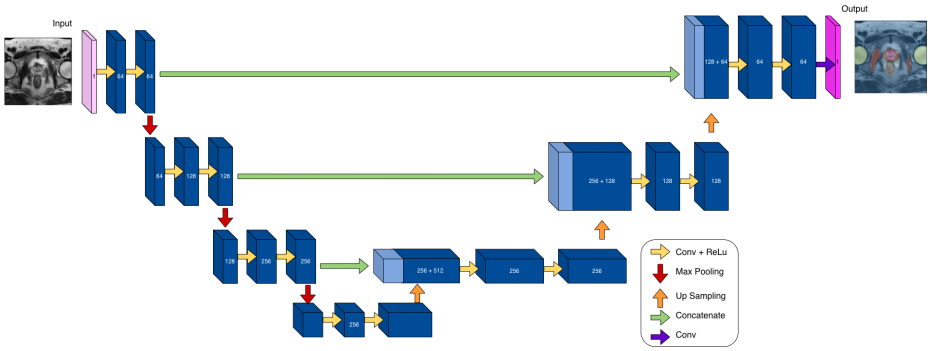

Fig. 1: A visualisation of the U-Net architecture.

2.1 Loss Function

The model was trained with cross entropy loss function for the flat case. A hierarchical loss function was derived to improve the model's performance using a penalty matrix applied to the cross entropy loss. Standard cross entropy treats all misclassifications equally, whereas a penalty matrix can increase the severity of certain misclassification over others.

The penalty matrix was designed to penalise misclassification based on the anatomical distance between misclassified structures, based on the work of Galdran et al. Specifically, more serious errors (e.g., biologically distant predictions such as the prediction of bone within the prostate) are assigned higher penalties, while less serious errors (e.g., incorrect predictions between neighbouring structures, such as the bladder vs the prostate) are assigned lower penalties [2].

A hyperparameter (α) was included in the loss function to tune how much the penalty matrix contributes to the overall loss. The final hierarchical loss function is as follows (with the full penalty matrix defined in the proceeding section).

$$\mathcal{L}_{hier} = \mathrm{CE}(y_i, \hat{y}_i) \cdot (1 + \alpha \cdot D_{y_i, \hat{y}_i}) \tag{1}$$

3 Experiments

To verify the performance of our hierarchical model, we conducted experiments with a U-Net trained with a hierarchical cost function, and a baseline U-Net trained with just cross entropy ($\alpha = 0$).

3.1 Datasets

We use the male pelvic MR dataset [3] for pre-training U-Net. This dataset includes 589 T2-weighted MRI's of the male pelvic region, with eight anatomical structures annotated. These structures include bladder, bone, obturator internus, transition zone, central gland, rectum, seminal vesicle and neurovascular bundle. For our application, the transition zone and central gland are grouped into the same level, as radiation therapy targets the prostate as a whole which includes these substructures [5, 6].

3.2 Implementation Details

All images and masks were first normalised and resampled to have a unified voxel spacing of $0.75 \times 0.75 \times 3$mm. To be robust to varying spatial resolutions, trilinear interpolation was applied to image volumes and nearest-neighbour was used. The dataset was split into training, validation and testing datasets with a 80:10:10 split. The PyTorch library was used and the model was trained on a GPU using CUDA. Models were trained for 50 epochs.

3.3 Penalty Matrix Definition

The penalty matrix $P \in \mathbb{R}^{9\times9}$ encodes the cost of misclassifying each anatomical structure. Diagonal entries and rows/columns corresponding to the background are set to zero, since correct classification and background misclassification are either not penalised or handled via cross-entropy. The matrix was generated by finding the average centroid location of each class label (excluding background) in the training mask dataset. The euclidean distance is then obtained between each class centroid to obtain the anatomical relationship of the class labels. To avoid centroids of symmetric structures (bone and the obturator internus) being outside of their structures, the anatomical distances were calculated separately for the left and right half of the image, after which the average was taken. To generalise; distances are rounded to the nearest 10 and normalised to $[0, 1]$. Prostate subclasses (Transition Zone and Central Gland) were merged to treat

prostate subclass misclassifications equally, and no penalties were assigned for intra-prostate misclassifications.

The resulting symmetric penalty matrix shown in Table 1 reflects the relative severity of misclassifications, with larger values for distant structures and smaller values for adjacent or intra-prostate classes. The detailed algorithm for its construction is shown in the code.

Table 1: Symmetric penalty matrix.

	0	1	2	3	4	5	6	7	8
0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1	0.00	0.00	0.91	0.64	0.45	0.45	0.82	0.64	0.55
2	0.00	0.91	0.00	0.45	0.82	0.82	1.00	0.73	0.73
3	0.00	0.64	0.45	0.00	0.36	0.36	0.55	0.27	0.27
4	0.00	0.45	0.82	0.36	0.00	0.00	0.36	0.18	0.09
5	0.00	0.45	0.82	0.36	0.00	0.00	0.36	0.18	0.09
6	0.00	0.82	1.00	0.55	0.36	0.36	0.00	0.27	0.36
7	0.00	0.64	0.73	0.27	0.18	0.18	0.27	0.00	0.09
8	0.00	0.55	0.73	0.27	0.09	0.09	0.36	0.09	0.00

3.4 Hyperparameter Tuning

The values of $[1, 3, 5, 10]$ were evaluated for α. The model was trained for 5 epochs for each hyperparameter value, with the best value selected based on the best recorded dice score. The model was reset between each new value tested. The final α chosen was 1, with Dice scores for each hyperparameter value being $[0.75, 0.73, 0.74, 0.74, 0.74]$.

3.5 Evaluation Metrics

Model evaluation was split for training/validation and testing. Training/validation evaluation was done through less computationally intensive methods: the loss function, dice score, and a superclass dice score that computes the dice score of the two prostate class predictions combined. The superclass dice captures the hierarchical objective of prostate segmentation, with overall prostate accuracy preferred over accuracy for prostate subclasses.

After training, the model with the best performance was evaluated with an additional metric: the hierarchical confusion matrix. It takes the non-normalized confusion matrix of the model and applies the penalty $(1 + \text{penalty\_matrix})$. This represents how the misclassifications change between models with special emphasis on more critical misclassifications, visualised as a heatmap in Figures 3 and 4.

4 Results

After training, models are evaluated on a held-out test set (n=60). Figure 2 shows the segmentation predictions of the both the flat and hierarchical model compared to the ground truth. The below sections detail and compare both the Flat Loss results and the Hierarchical Loss results.

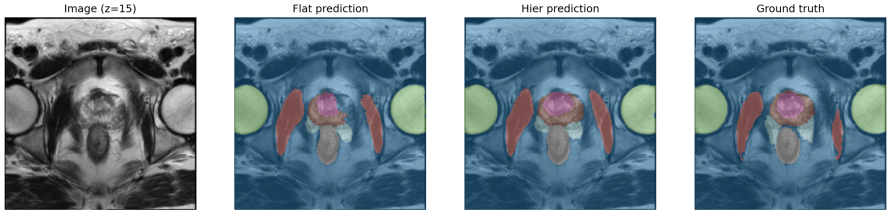

Fig. 2: Overlay of Flat and Hierarchical Best Model Predictions Image 1006 [3]

4.1 Flat Loss Results

Table 2 shows that large structures (bladder, bone, rectum) achieved Dice $>$ 0.83, while the neurovascular bundle proved most challenging (0.512 $\pm$ 0.169). The prostate superclass achieved 0.832 $\pm$ 0.084.

4.2 Hierarchical Loss Results

The hierarchical model ($\alpha = 1$) achieved mean foreground Dice of 0.765. Table 2 presents per-class results with changes relative to baseline.

Five of eight structures improved under hierarchical training, with the largest gains for neurovascular bundle (+3.5%) and central gland (+1.7%). Prostate superclass dice improved to 0.838 with reduced variance (std: 0.072 vs. 0.084).

Table 2: Per-class Dice scores for flat and hierarchical models.

Structure	**Flat Model**	**Hier Model**	Δ
	Dice $\pm$ std	**Dice Score $\pm$ std**	
Bladder	0.880 $\pm$ 0.135	0.872 $\pm$ 0.149	-0.8%
Bone	0.883 $\pm$ 0.150	**0.892 $\pm$ 0.133**	+1.0%
Obturator Internus	0.847 $\pm$ 0.052	0.845 $\pm$ 0.051	-0.2%
Transition Zone	0.684 $\pm$ 0.104	0.687 $\pm$ 0.100	+0.3%
Central Gland	0.789 $\pm$ 0.119	**0.802 $\pm$ 0.092**	+1.7%
Rectum	0.838 $\pm$ 0.091	0.829 $\pm$ 0.084	-1.1%
Seminal Vesicle	0.661 $\pm$ 0.141	0.665 $\pm$ 0.132	+0.6%
Neurovascular Bundle	0.512 $\pm$ 0.169	**0.530 $\pm$ 0.157**	+3.5%
Prostate (Superclass)	**0.832 $\pm$ 0.084**	**0.838 $\pm$ 0.072**	+0.7%

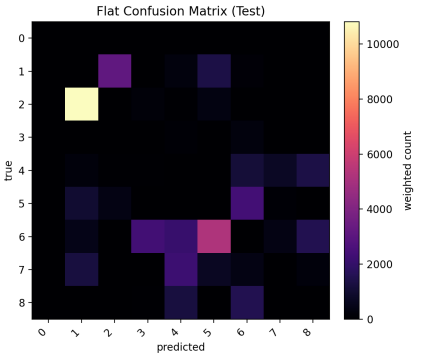

Fig. 3: Flat loss model hierarchical confusion matrix.

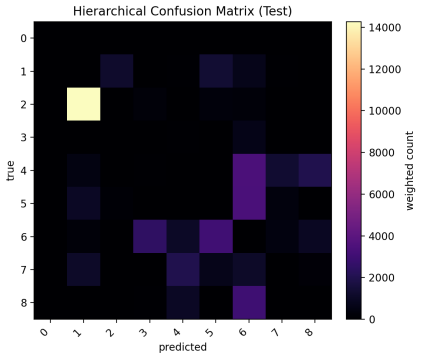

Fig. 4: Hierarchical loss model hierarchical confusion matrix.

5 Discussion and Conclusions

The results indicate that the hierarchical loss model offers a slight improvement in the classification of the prostate, but not significant. The hierarchical loss approach led to improvements in classification for other organs such as the neurovascular bundle. Visually, the hierarchical loss model shows an improved segmentation of the prostate (Figure 2), but this observation cannot be generalised to all image slice samples.

The confusion matrices (Figures 3 and 4) reveal that the hierarchical loss model reduced misclassifications between the prostate and distant organs such as the bladder, but increased misclassifications of the prostate with adjacent organs such as the seminal vesicle.

A revealing result, discussed in hyperparameter tuning, indicates that the hyperparameter has negligible impact on the dice score despite having major impacts on the loss function at higher values. The penalty matrix therefore has no real impact on the dice score regardless of its magnitude. This could be attributed to the matrix values being relatively close together; the penalty matrix could be reconstructed to take squared or exponential distances between organ, or otherwise be simplified to explicitly penalise prostate misclassifications instead of misclassifications between all classes.

The hierarchical loss method appears to show promise in improving prostate segmentation ability. Future research can focus on testing different penalty matrices to address errors such as misclassifications between the prostate and the adjacent seminal vesicle. Alternative distance formulations when constructing the penalty matrix are also an avenue for possible enhancement in segmentation performance. With these refinements the hierarchical loss function may become more suitable for prostate segmentation to support clinicians during pelvic radiotherapy procedures.

References

1. El-Baz, A., Gimel'Farb, G., Suzuki, K.: Machine Learning Applications in Medical Image Analysis. Computational and Mathematical Methods in Medicine **2017**, 2361061 (2017). https://doi.org/10.1155/2017/2361061, https://pmc.ncbi.nlm.nih.gov/articles/PMC5406720/
2. Galdran, A., Dolz, J., Chakor, H., Lombaert, H., Ben Ayed, I.: Cost-Sensitive Regularization for Diabetic Retinopathy Grading from Eye Fundus Images
3. Li, Y., Fu, Y., Gayo, I.J., Yang, Q., Min, Z., Saeed, S.U., Yan, W., Wang, Y., Noble, J.A., Emberton, M., Clarkson, M.J., Huisman, H., Barratt, D.C., Prisacariu, V.A., Hu, Y.: Cross-institution Male Pelvic Structures [Data set] . https://doi.org/10.5281/ZENODO.7013610, https://zenodo.org/records/7013610
4. Muller, B.R., Smith, W.A.P.: A Hierarchical Loss for Semantic Segmentation https://orcid.org/0000-0003-3682-9032
5. Murgić, J., Gregov, M., Mrčela, I., Budanec, M., Krengli, M., Fröbe, A., Franco, P.: MRI-GUIDED RADIOTHERAPY FOR PROSTATE CANCER: A NEW PARADIGM. Acta Clinica Croatica **61**(Suppl 3), 65 (2022). https://doi.org/10.20471/ACC.2022.61.S3.9, https://pmc.ncbi.nlm.nih.gov/articles/PMC10022406/
6. Pasquier, D., Lacornerie, T., Vermandel, M., Rousseau, J., Lartigau, E., Betrouni, N.: Automatic Segmentation of Pelvic Structures From Magnetic Resonance Images for Prostate Cancer Radiotherapy. International Journal of Radiation Oncology*Biology*Physics **68**(2), 592–600 (6 2007). https://doi.org/10.1016/J.IJROBP.2007.02.005
7. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **9351**, 234–241 (2015). https://doi.org/10.1007/978-3-319-24574-4{_}28, https://link.springer.com/chapter/10.1007/978-3-319-24574-4$_2$8