Oropharyngeal Tumour Segmentation Using Ensemble 3D PET-CT Fusion Networks for the HECKTOR Challenge

Oropharyngeal Tumour Segmentation Using Ensemble 3D PET-CT Fusion Networks for the HECKTOR Challenge

Chinmay Rao[1(✉)], Suraj Pai[1], Ibrahim Hadzic[1], Ivan Zhovannik[1,2], Dennis Bontempi[1], Andre Dekker[1], Jonas Teuwen[3], and Alberto Traverso[1]

[1] Department of Radiation Oncology (Maastro), GROW School for Oncology, Maastricht University Medical Centre+, Maastricht, The Netherlands
chinmay.rao@maastro.nl
[2] Department of Radiation Oncology, Radboud Institute of Health Sciences, Radboud University Medical Centre, Nijmegen, The Netherlands
[3] Department of Medical Imaging, Radboud University Medical Centre, Nijmegen, The Netherlands

Abstract. Automatic segmentation of tumours and organs at risk can function as a useful support tool in radiotherapy treatment planning as well as for validating radiomics studies on larger cohorts. In this paper, we developed robust automatic segmentation methods for the delineation of gross tumour volumes (GTVs) from planning Computed Tomography (CT) and FDG-Positron Emission Tomography (PET) images of head and neck cancer patients. The data was supplied as part of the MICCAI 2020 HECKTOR challenge. We developed two main volumetric approaches: A) an end-to-end volumetric approach and B) a slice-by-slice prediction approach that integrates 3D context around the slice of interest. We exploited differences in the representations provided by these two approaches by ensembling them, obtaining a Dice score of 66.9% on the held out validation set. On an external and independent test set, a final Dice score of 58.7% was achieved.

Keywords: Oropharyngeal cancer · Radiotherapy treatment planning · Automatic segmentation · Multi-modal · PET-CT · 3D U-Net

1 Introduction

According to the European Society for Medical Oncology, Head and Neck Squamous Cell Carcinoma (HNSCC) is the sixth most frequently occurring cancer globally [7]. Among all cancer types, HNSCC accounts for 6% of the occurrences and 1–2% of the deaths. Radiotherapy is standard of care treatment for HNSCC.

C. Rao and S. Pai—Equal contribution.

J. Teuwen and A. Traverso—These authors share senior authorship.

V. Andrearczyk et al. (Eds.): HECKTOR 2020, LNCS 12603, pp. 65–77, 2021.
https://doi.org/10.1007/978-3-030-67194-5_8

PET-CT scans are usually employed for treatment planning. The treatment planning process involves manual contouring of the gross tumour volumes (GTV) which is time-consuming, expensive, and suffers from inter- and intra-reader variability. Accurate and robust automatic segmentation can potentially solve these issues. In addition to treatment planning, the field of radiomics [1] can also benefit from reliable automatic segmentation algorithms for PET-CT images. Radiomics involves predicting tumour characteristics using image-derived quantitative biomarkers. Large scale validation of PET-CT based radiomics models is currently limited by a shortage of PET-CT image datasets containing precise expert-delineated GTVs, and can be tackled by applying automatic segmentation techniques to generate GTV segmentation from unlabelled data. This is the primary motivation behind the inception of the MICCAI 2020: HECKTOR challenge [2,3] which seeks to evaluate bi-modal fusion approaches for segmentation of oropharyngeal GTV in FDG-PET/CT volumes.

To handle the complementary bi-modal information, a variety of approaches have been proposed in recent literature. Andrearczyk et al. [2] employ two simple PET-CT fusion strategies, namely early fusion and late fusion, in a V-Net based framework for segmenting head-and-neck GTV and metastatic lymph nodes. Zhong et al. [18] apply a late fusion approach using two independent 3D U-Nets for PET and CT respectively, and graph-cut co-segmentation to combine their outputs. Novel and specialised deep neural architectures which incorporate fusion of PET and CT-derived information have also been proposed, for example, a two-stream chained architecture [9], a specialised W-Net architecture for bone-lesion detection [16], multi-branched networks that seek to fuse deep features learnt separately from PET and CT and then co-learn the combined features [10,17], and a modular architecture using multi-modal spatial attention [8]. Li et al. [11] propose a hybrid approach that utilises a 3D fully convolutional network to obtain tumour probability map from CT and fuse it with PET data using a fuzzy variational model.

In this paper, we describe an ensemble based segmentation model consisting of two 3D U-Net based networks, and we compare various strategies for combining their outputs based on volumetric Dice score. We explore simple ensembling methods including weighted averaging, union, and intersection operations. We discuss in detail our segmentation approach in Sect. 2.3 and Sect. 2.4. Additionally, we compare commonly used pre-processing methods and investigate the effects of post-processing on the model performance. Details of the pre-processing and post-processing schemes are described in Sect. 2.2 and Sect. 2.5, respectively. The experiments performed for the aforementioned comparison studies as well as their results are documented in Sect. 3. Finally, we make the code for most of the data operations performed publicly available.[1]

[1] https://gitlab.com/UM-CDS/projects/image-standardization-and-domain-adaptation/hecktor-segmentation-challenge.

2 Methodology

2.1 Dataset

For training and validating our models, we used the benchmark dataset supplied for the HECKTOR challenge. The training set consists of FDG-PET/CT data and the corresponding GTV segmentation masks from 201 patients diagnosed with oropharyngeal cancer obtained from four centres in Québec (Canada). This corpus of data is a subset of a larger dataset originally proposed by Vallières et al. [15], which is publicly available on The Cancer Imaging Archive [6,12]. For the purpose of the challenge, this subset underwent quality control, including the conversion of raw PET intensties to SUV and the reannotation of the primary GTV for each patient. The test dataset, provided for the HECKTOR challenge and used for the final evaluation of the submissions, is a set of FDG-PET/CT scans from 53 patients from the Centre Hospitalier Universitaire Vaudois (Lausanne, Switzerland). The supplied imaging data vary in physical size, array size and voxel spacing across patients. Hence, for the purpose of standardisation, we cropped all the images to $144 \times 144 \times 144\,\text{mm}^3$ physical size using simple PET-based brain segmentation. Subsequently, we resampled the scans using 3^{rd} order spline interpolation to have a pixel spacing of $1 \times 1\,\text{mm}^2$ in the x-y plane and 3 mm spacing between axial slices. These dimensions were chosen by obtaining a distribution of pixel spacing across the entire dataset and choosing the mode of the distribution to minimise oversampling in comparison to isotropic resampling. The supplied HECKTOR training set was randomly split to produce two subsets with 180 and 21 patients for model training and validation respectively. The aforementioned data preparation steps were implemented using code obtained from the public Github repository released by the challenge organisers[2].

2.2 Pre-processing

The voxel intensities of the resampled CT and PET modalities are measured in Hounsfield Units (HU) and Standardised Uptake Values (SUV), respectively. In the supplied dataset, the PET scan intensities were already converted from absolute activity concentration (Bq/mL) and counts (CNTS) units to SUV. We processed the HU values by applying a window between the range $[-150, 150]$ to focus on tissues within the particular range, which include the GTV. We subsequently normalised this to a range of $[0, 1]$. The maximal SUV values are more dynamic in range compared to the HU values, although between a range of $[0, 5]$ the values follow similar distributions across the dataset. This behaviour of similar distributions between $[0, 5]$ can be seen in Fig. 1. In order to account for this, we limited the SUV values between the range $[0.01, 8]$ following which a $[0, 1]$ normalisation was performed. We refrained from using global normalisation schemes to avoid value shift on the test data which was collected from a different centre. As an alternate normalisation scheme, z-score normalisation was

[2] http://github.com/voreille/hecktor.

also explored for the PET-CT pair but min-max [0, 1] normalisation ultimately provided the best performance on our validation data.

2.3 Network Architectures

In this study, we test two different network architectures, comparing the strengths and drawbacks of their associated input-output representations. The first network architecture infers the segmentation masks on a slice-by-slice basis, by integrating "a slice context" around each slice to be predicted. At each of these slices, the network outputs predictions based on values of the slice and the values of its neighbours in a certain range. This range of neighbours around a particular slice is what we term as slice context. The second network is a fully volumetric 3D network that takes a full volume as input and outputs another volume containing predictions for each voxel of the input.

3D-to-2D U-Net with Fully Connected Bottleneck. We used a custom implementation of the 3D U-Net network architecture proposed by Nikolov et al. [13]. The input to this network is a 3D volume with 21 slices with a dimension of 128 × 128 each. Of the 21 slices, 10 slices at each side of the central slice comprise the slice context, and the network outputs a 2D segmentation corresponding to the central slice. In terms of network design, 7 down-convolutional blocks with a mix of 2D and 3D convolutions are present in the analysis path of the U-Net. At the end of the analysis path, a fully connected bottleneck is introduced. Following the bottleneck, 7 up-convolutional blocks give way to the synthesis path of the network.

Fully Volumetric 3D U-Net. The 3D U-Net was introduced by Çiçek et al. [5] to extend the success of 2D U-Nets to 3D volumetric inputs. In our work, we used a modified version of the 3D U-Net provided as part of the ELEKTRONN3 Toolkit[3]. The input to this network is a 3D volume with 48 slices in the z axis and each slice has a shape of 144 × 144. The output of the network follows the same spatial configuration as the input. A shallow architecture is used in order to allow fitting more 3D volumes per batch. 3 down-convolutional blocks are used in the analysis path and give way to 2 up-convolutional blocks in the synthesis path.

2.4 Model Training and Hyperparameters

A large amount of focus in our work was placed on model training procedures to account for 3D data, data imbalance, and memory and computational efficiency. For the network described in Sect. 2.3, we used label-based sampling where slices were selected by sampling randomly from all the slices that contain the GTV. In order to account for slices where the GTV is absent we also randomly sampled

[3] http://elektronn3.readthedocs.io.

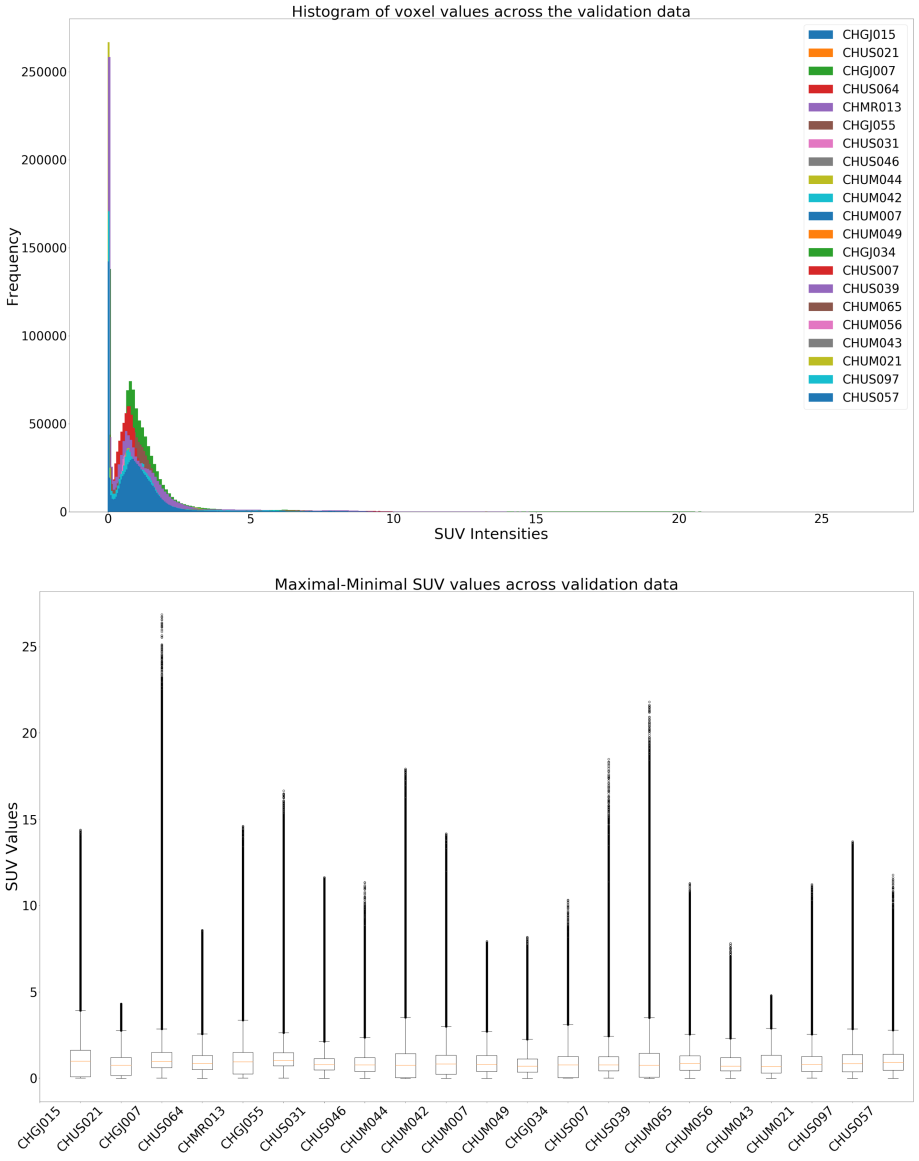


Fig. 1. SUV values of the validation data with a histogram and box plot to show the distribution of voxel intensities and the maximal and minimal values. The legend in figure (a) and the x-axes labels in (b) correspond to subject IDs in the validation data. The first 4 alphabets correspond to a centre and the 3 digits following correspond to a numeric ID. For example, CHUM007 corresponds to a subject 7 from Centre Hospitalier de l'Université de Montréal

from background slices with a certain probability (chosen to be $p = 0.2$). Data imbalance was tackled by using a top-k loss similar to [13] which optimises losses from $k\%$ of worst performing voxels in an image. This allows the model to deal efficiently with imbalanced data points and to train the hardest losses first. The optimiser was attached to a decaying cyclic learning rate scheduler [14] to help it deal with the complex loss landscape obtained through the top-k loss. By using a range of learning rates, the scheduler ensures that the optimiser can jump out of local minima and avoids stagnation during training compared to decaying learning rate schedulers.

2.5 Post-processing

We employed a post-processing step to refine the predicted GTV structure in the model's hypothesised binary masks as well as to address false positive voxel groups. First, morphological dilation was performed using a $5 \times 5 \times 5$ structuring matrix with a roughly spherical structure in order to make the predicted structure more globular as tumours generally are. Following this, all connected components from the binary image were extracted and the largest geometrical structure was considered the GTV while disregarding all the others as false positives. Finally, a morphological closing was applied on the largest connected component to smooth the contours with a similar structuring matrix as the dilation.

3 Experiments and Results

In this section, we describe the experiments performed and consequently the results obtained using our methodologies. All experiments were tracked using Weights and Biases (W&B) [4] to observe qualitative metrics such as per scan predicted segmentation maps and quantitative metrics such as loss and Dice scores. We provide the W&B run info corresponding to each of our experiments to allow reproducibility, hosted on this dashboard.

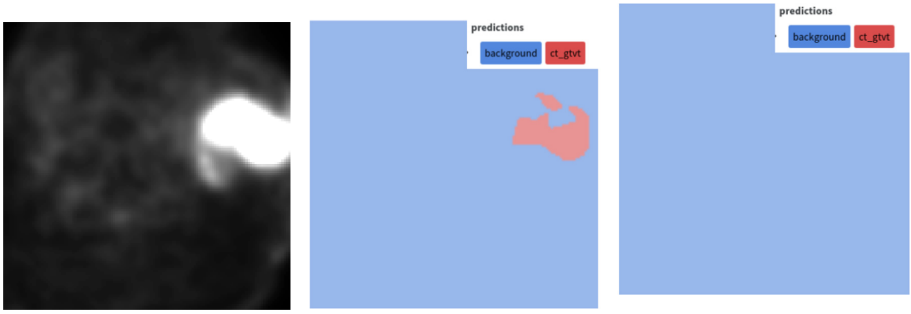
All our experiments were run on clusters provided by the Data Science Research Infrastructure at Maastricht University[4] and the HPC cluster hosted by RWTH-Aachen[5]. Due to differences in hardware across these clusters, we used different batch sizes (32 for the 3D-to-2D U-Net and 8 for the fully volumetric 3D U-Net) and caching methodologies to perform efficient training. These details can be found by exploring the W&B dashboard.

3.1 PET only Training

As a preliminary experiment, the network in Sect. 2.3 was trained only on PET data. After training for 200,000 iterations[6] using the training configurations mentioned in Sect. 2.4, with a learning rate range of 0.001 to 0.01, a final Dice score

[4] https://maastrichtu-ids.github.io/dsri-documentation/.
[5] https://doc.itc.rwth-aachen.de/.
[6] Each iteration corresponds to one forward-backward pass over a batch.

of 0.526 was obtained on the held out validation data. Qualitatively inspecting the obtained results showed that there were numerous cases where high probability of GTV was seen when the PET intensity was high but there was no tumour present in the ground truth. Figure 2 shows an example of the false positives seen. Pairing this PET data with structural information would play a strong role in avoiding such cases and discriminating between false positives in high intensity regions.



(a) Input Image **(b)** Predicted Label Map **(c)** Ground Truth Label Map

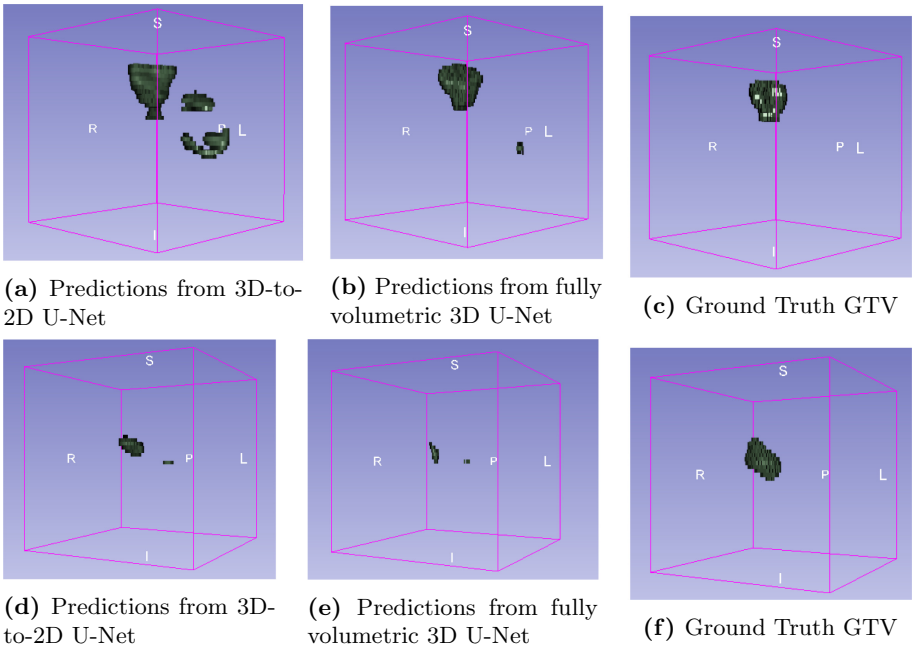
Fig. 2. False positives predicted by the PET-only network when high intensity regions are seen in the validation set. The legend of semantic labels can be seen on to the top right corner of the label map images. *ct_gtvt* label map corresponds to the presence of tumour within that region while *background* corresponds to its absence.

3.2 PET-CT Early Fusion

To fuse information from both the PET and CT modalities, we applied a very straightforward channel-wise fusion strategy. The PET and CT 3D volumes were stacked across channels forming a 4D PET-CT input to the models. We followed this fusion strategy to allow the entire model to have access to combined PET-CT information as they are complementary in determining GTV contours.

The PET-CT data was fed as input into both the networks defined in Sect. 2.3. For the 3D-to-2D model described in Sect. 2.3, we used a large batch size owing to the smaller 3D input in comparison to the end-to-end volumetric 3D approach. Both networks were also run for 200,000 iterations with these batch sizes. The 3D-to-2D network was trained with rotate, shear and elastic deformations applied on the fly during training time. The results of the training led to qualitatively and quantitatively superior results compared to the PET-only network achieving a Dice score of 0.648 on the validation data split. Qualitatively the results also show increased true positives and a huge decrease in false positives that spiked with higher intensity values as seen in the PET-only approach. A fully volumetric 3D approach was also experimented with—to compare

against the 3D volume to 2D slice prediction input-output representation. This experiment provided quantitative results similar to the former approach with a Dice score of 0.639 but differed in the qualitative predictions. The qualitative predictions of this network were significantly smoother in the 3D space than the previous approach but some smaller contours (occupying smaller dimensions in the voxel space) were missed. Figure 3 shows these qualitative differences across different networks.



(a) Predictions from 3D-to-2D U-Net

(b) Predictions from fully volumetric 3D U-Net

(c) Ground Truth GTV

(d) Predictions from 3D-to-2D U-Net

(e) Predictions from fully volumetric 3D U-Net

(f) Ground Truth GTV

Fig. 3. GTV predictions for the two types of models compared with the ground truth in 3D. For the images in the first row, (b) has significantly fewer false positives compared to (a) and produces more accurate 3D volumes. In the second row, (d) trumps (e) in terms of correspondence to the ground truth. (e) misses out largely in matching the shape of the contour in (f)

3.3 Model Ensembling

From visual inspection of a subset of predictions from both 3D-to-2D and fully volumetric 3D early fusion U-Nets, we found that each of these networks failed in ways different from the other in correctly segmenting the GTV. This can be seen in Fig. 3. In order to utilise this apparent complementary behaviour, we experimented combining their outputs using simple ensembling approaches to produce the final segmentation mask for each of the validation examples. In

particular, we compared three operations - weighted voxel-wise average, union and intersection. The weighted average operation was applied to the output voxel-wise probabilities of the two networks where a single fixed weight was assigned to the output probability map of each network. Union and intersection operations were applied to the binary mask outputs of the two networks each obtained using a GTV probability threshold of 0.5.

3.4 Post-processing

To measure the effect of the post-processing sequence discussed in Sect. 2.5 on the model performance, the post-processing operations were applied to the binary prediction masks of models in every case - the two early-fusion U-Net models and their ensembles - and the resulting validation Dice scores were compared with those of the corresponding models without the post-processing step.

Weight values used for the weighted average ensembling operation were 0.6 and 0.4 for 3D-to-2D and fully volumetric 3D U-Nets respectively when no post-processing was performed. With post-processing, the weights were 0.5 for both. In each case, the weights were optimal among a fixed set of values with respect to the average validation Dice, as shown in Fig. 4.

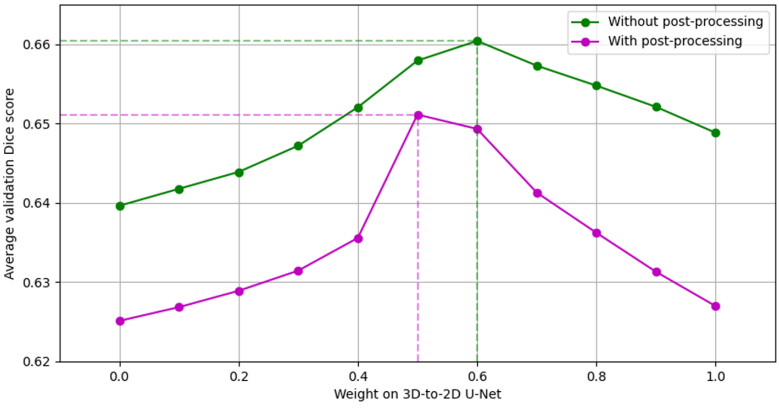


Fig. 4. Weight vs. average validation Dice plot for ensemble model with weighted voxel-wise average strategy with and without the post-processing step. Weight, here, refers to the weight value assigned to the predicted GTV probability map of 3D-to-2D U-Net.

Table 1 shows the validation Dice score for each of the aforementioned model configurations we experimented with. Combining the predictions of the two early fusion U-Nets with weighted voxel-wise average improved the overall performance. This improvement was also observed when post-processing was introduced. The use of post-processing step, however, deteriorated the final score in every case rather than improving it, except for the intersection based ensemble which exhibited a slight improvement. The negative effect of post-processing

can be observed on the union ensemble's performance as it sharply dropped from being the best among the others to being the worst. Moreover, in the case of weighted average ensemble, post-processing results in a deteriorated performance for all weights as seen in Fig. 4.

Table 1. Average validation Dice score (DSC) for each of the model and ensemble variants tested.

Network variants	DSC w/o Post-processing	DSC with Post-processing
3D-to-2D U-Net	0.648	0.626
Fully volumetric 3D U-Net	0.639	0.625
Intersection	0.614	0.635
Union	0.669	0.618
Weighted average	0.657	0.651

A block diagram overview of the different components in the segmentation pipeline can be found in Fig. 5. Individual components of the pipeline were described in detail in the preceding sections.

3.5 Post-challenge Results

The best performing model variant on the held out test-set from the challenge was the weighted average ensemble without post-processing applied. This is seen in Fig. 4 at the peak of the green line plot. The 3D-to-2D U-Net predictions, p_1 and the fully volumetric 3D U-Net predictions, p_2 are combined as,

$$p = 0.6 \times p_1 + 0.4 \times p_2 \tag{1}$$

The final binary label map, obtained by thresholding $p >= 0.5$, is submitted to the challenge. With this, we obtain a Dice score of 0.587. Compared to the challenge winners, we see a large drop in our Dice scores (-0.17). We hypothesise this difference to be due to distribution-shift across data from different centres and our method's inability to account for these in the data pre-processing strategies.

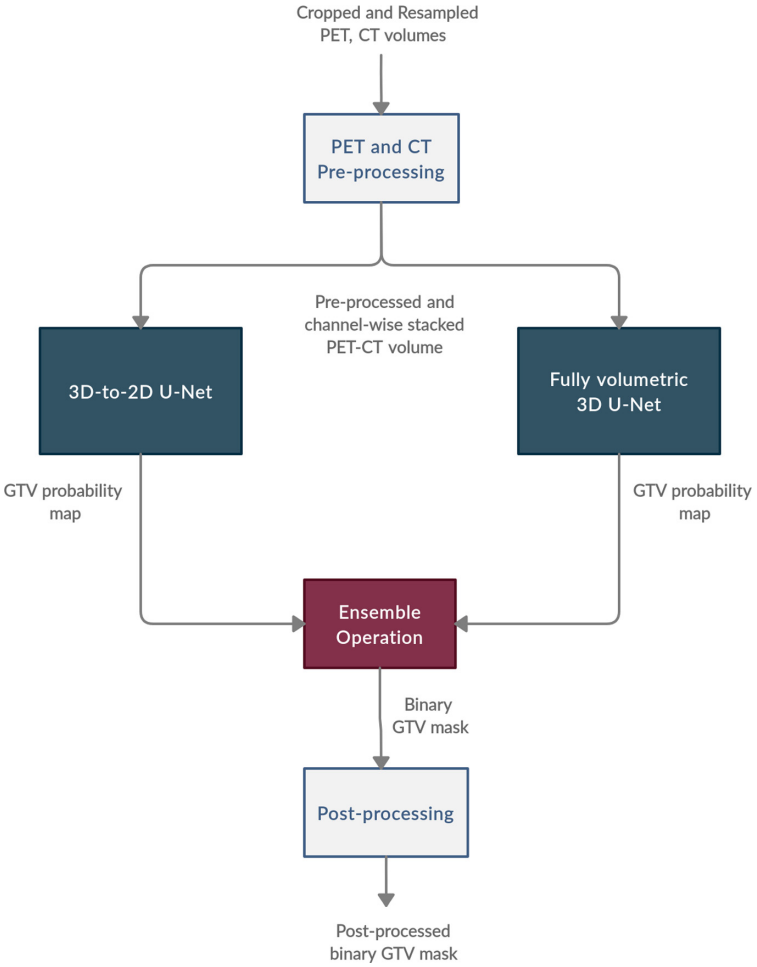


Fig. 5. Block diagram of the segmentation pipeline. The entire procedure followed in obtaining a predicted GTV mask from the PET-CT dataset provided as a part of the challenge is presented in the diagram.

4 Discussion and Conclusion

The HECKTOR challenge provides a strong benchmark to compare automatic segmentation methods for oropharayngeal tumours in PET-CT images. Development of automatic methods can prove to be highly useful in providing delineation assistance in radiotherapy treatment planning as well as for advancement of PET-CT based radiomics by facilitating the generation of segmentation data for validation of radiomics methods on large cohorts. Through the challenge, we were able to compare different 3D approaches with varied input-output representations, pre-processing methods and training hyperparameter schemes.

A stark difference was observed between PET-CT fusion and PET-only network results which quantitatively bolsters the importance of complementary information provided by the combined modalities. After obtaining slice-by-slice prediction models and fully volumetric 3D prediction models, ensemble methods were investigated to combine strengths across these methods.

The study performed by Andrearczyk et al. [2] includes using the early-fusion strategy in a 3D V-Net architecture, among other design choices and combinations, to segment primary orophayngeal GTV and metastatic lymph lodes. Although a meaningful comparison of our results with theirs cannot be performed due to differences in the data used, it would be interesting to study the influence of architecture design on the model performance. For instance, a comparison between the 3D V-Net and the fully volumetric 3D U-Net design used in this study, in the context of PET-CT early-fusion.

As future work, we plan to conduct larger cross validation studies across centres to enable a meaningful comparison with other approaches. Additionally, cross validation strategies that can account for distribution-shift can help us improve generalisation ability of our methods to new test centres as seen in the held out test-set. To shed light on the results in a qualitative manner and to incorporate clinicians into the process, a Turing test could be performed to analyse how satisfied a radiation oncologist would be with the tumours automatically delineated by our methods.

References

1. Aerts, H.J., et al.: Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat. Commun. **5**(1), 1–9 (2014)
2. Andrearczyk, V., et al.: Automatic segmentation of head and neck tumors and nodal metastases in PET-CT scans. In: International Conference on Medical Imaging with Deep Learning (MIDL) (2020)
3. Andrearczyk, V., et al.: Overview of the HECKTOR challenge at MICCAI 2020: automatic head and neck tumor segmentation in PET/CT. In: Andrearczyk, V., et al. (eds.) HECKTOR 2020. LNCS, vol. 12603, pp. 1–21. Springer, Cham (2021)
4. Biewald, L.: Experiment tracking with weights and biases (2020). https://www.wandb.com/. Software available from wandb.com
5. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 424–432. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_49
6. Clark, K., et al.: The cancer imaging archive (TCIA): maintaining and operating a public information repository. J. Digit. Imaging **26**(6), 1045–1057 (2013). https://doi.org/10.1007/s10278-013-9622-7
7. Economopoulou, P., Psyrri, A.: Head and Neck Cancers: Essentials for Clinicians, chap. 1. ESMO Educational Publications Working Group (2017)
8. Fu, X., Bi, L., Kumar, A., Fulham, M., Kim, J.: Multimodal spatial attention module for targeting multimodal PET-CT lung tumor segmentation. arXiv preprint arXiv:2007.14728 (2020)

9. Jin, D., et al.: Accurate esophageal gross tumor volume segmentation in PET/CT using two-stream chained 3D deep network fusion. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11765, pp. 182–191. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32245-8_21
10. Kumar, A., Fulham, M., Feng, D., Kim, J.: Co-learning feature fusion maps from PET-CT images of lung cancer. IEEE Trans. Med. Imaging **39**(1), 204–217 (2019)
11. Li, L., Zhao, X., Lu, W., Tan, S.: Deep learning for variational multimodality tumor segmentation in PET/CT. Neurocomputing **392**, 277–295 (2020)
12. Martin, V., et al.: Data from head-neck-PET-CT. The Cancer Imaging Archive (2017)
13. Nikolov, S., et al.: Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy. CoRR abs/1809.04430 (2018). http://arxiv.org/abs/1809.04430
14. Smith, L.N.: No more pesky learning rate guessing games. CoRR abs/1506.01186 (2015). http://arxiv.org/abs/1506.01186
15. Vallieres, M., et al.: Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer. Sci. Rep. **7**(1), 1–14 (2017)
16. Xu, L., et al.: Automated whole-body bone lesion detection for multiple myeloma on 68ga-pentixafor PET/CT imaging using deep learning methods. Contrast Media Mol. Imaging **2018** (2018)
17. Zhao, X., Li, L., Lu, W., Tan, S.: Tumor co-segmentation in PET/CT using multimodality fully convolutional neural network. Phys. Med. Biol. **64**(1), 015011 (2018)
18. Zhong, Z., et al.: 3D fully convolutional networks for co-segmentation of tumors on PET-CT images. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pp. 228–231. IEEE (2018)