Physics and Imaging in Radiation Oncology 23 (2022) 150–152

Contents lists available at ScienceDirect

Physics and Imaging in Radiation Oncology

journal homepage: www.sciencedirect.com/journal/physics-and-imaging-in-radiation-oncology

Editorial

# Challenges and chances for deep-learning based target and organ at risk segmentation in radiotherapy of head and neck cancer

One of the most important steps in the radiotherapy treatment chain is the definition of the target volume and the organs at risk. Especially in head and neck cancer (HNC) patients this can be a tedious task, as many organs at risk (OARs) are present, and primary tumor (GTVt), involved lymph nodes (GTVn), and elective target volumes need to be segmented. With the rapid development of deep-learning based medical image segmentation (DLS) in the past decade, research regarding auto-segmentation of OARs and target volumes in HNC is rapidly increasing. This trend was kickstarted with the 2015 HNC auto-segmentation challenge for OARs [1], and the 2020 HNC tumor segmentation challenge (HECKTOR) [2]. In these challenges, research teams from all over the globe compete in achieving the most accurate segmentations on a publicly provided dataset. These public datasets remain impactful beyond the scope of the challenges, as they are often used to demonstrate the impact of new DLS developments, as training set, or external validation set. This enables comparison of research results, but also limits the variability of data-quality used in research leading to a potential mismatch in segmentation accuracy with real-world data.

Themes that often are addressed in DLS research evolve around the impact of dataset size, dealing with the class imbalance problem, and especially for GTV segmentation, uni- vs multi-modal imaging. Recently, two papers in the area of DLS in HNC were published in *Physics and Imaging in Radiation Oncology*, by Henderson et al. [3] and Outeiral et al. [4]. The Henderson et al. paper focused on OAR segmentation (brainstem, mandible, parotid glands, spinal cord) on computed tomography (CT) scans, while the latter addressed primary tumor segmentation in the oropharynx on magnetic resonance imaging (MRI)-only data.

## 1. OAR segmentation in HNC

Henderson et al. [3] used a publicly available CT dataset of 35 patients [5], and assessed the influence of using one or three input channels with different window level settings (soft-tissue, bone, brain), three different loss functions (multi-class weighted soft-dice (wSD), cross-entropy (XE) + wSD, and Exponential Logarithmic Loss (ExpLogLoss), and the use of transpose vs resize convolutions in the up-sampling part of their convolutional neural network (CNN). For external validation, they also took their optimal model configuration and trained with the 2015 OAR challenge dataset, using the 25 patients of the training set for training, the 5 onsite testing patients for validation, and the 10 offsite patients for testing. The main findings were that the ExpLogLoss was the best loss function, and using three input channels with different window level settings improved results for the soft-tissue OARs, but not for the mandible. For the parotids, the three input channels $HD_{95\%}$ scores were about 1 mm smaller compared to one channel. The type of convolutions had limited effect on the segmentation accuracy. The segmentation results on the 2015 OAR challenge dataset were comparable to the state-of-the-art papers on the same data.

It is interesting to see the impact of a simple pre-processing step offering one CT scan in three different contrast settings to the CNN. Almost at the same time as the Henderson et al. paper, a study on class imbalance in HN OAR segmentations by Tappeiner et al. got published [6]. They tested two strategies on the 2015 OAR challenge dataset using the nnUNet framework [7]. The first was to optimize the patch size to minimize the class imbalance. The second was to adapt the Dice loss function by making it class-adaptive, meaning that only classes available in the patch are in the calculation, instead of assuming a perfect score for missing classes. The main improvement came from reducing the standard patch size (192 * 160 * 56) to a smaller patch size (90 * 80 * 48), resulting in a Dice similarity coefficient (DSC) increase of 0.02–0.03 and a reduction of the 95 % Hausdorff distance ($HD_{95\%}$) from 4 to 3 mm. The adaptation of the loss function had very limited effect.

When comparing Henderson et al. to Tappeiner et al. there is an interesting difference. Henderson et al. did not use patches in their training, but automatically cropped the scans to anatomically consistent sub-volumes with the dimensions of 200 × 200 × 56 voxels [8], slightly larger than the standard patch size in the Tappeiner et al. paper. This automated cropping step assured that only the relevant part of the CT scans was used for training, and possibly had a bigger impact on the segmentation accuracy than the presented methods. It would be interesting to know if patch size would still affect results when imaging data is first automatically cropped to the right area of interest.

In the ideal world, testing of methods that optimize CNN performance in the setting of limited training data need to be evaluated in a large dataset, to put performance into perspective. Fang et al. nicely showed that for most HNC OARs, a training set of 40 patients can produce DSC scores of about 95 % of what you are able to reach if you have a larger dataset of up to 800 patients [9]. Unfortunately, DSC is not the full story of segmentation accuracy, as it is volume dependent, and MSD and $HD_{95\%}$ might be more representative for how segmentation differences influence the dose distribution. One can speculate that for MSD and especially for $HD_{95\%}$, more patients are needed to achieve similar levels, as these measures are more influenced by inter-observer variation in the ground truth data. Unfortunately, both of these measures were not presented in the study by Fang et al. [9].

Disentangling the problem of inter-observer variation (IOV) from the accuracy of DLS tools is challenging, as the manual contour is the only

https://doi.org/10.1016/j.phro.2022.08.003

Available online 11 August 2022

"truth" available, and it is dependent on who has done the segmentation. With more and more DLS tools for OARs being used in the clinics, it now becomes possible to assess auto-segmentation accuracies in clinical practice [10,11]. By assessing the adaptations of the auto-contours, insight is obtained into where observers disagree with the DLS tools. Brouwer et al. [10] showed in a study on 103 HNC patients that organs at risk can be put in a standardized reference frame, to systematically study where DLS contours are being adapted. They showed that some of the systematic adaptations could be linked to the performance of the DLS tool, while others were based on interpretation differences of delineation guidelines between observers. A similar study on OAR segmentation in lung (n = 350) and breast cancer patients (n = 362) showed that analyzing adaptations of auto-contours from DLS in real world data is a powerful tool to identify potential improvements in the DLS tools, the workflow, and the scan protocols.

## 2. GTV segmentation in HNC

Use of DLS tools in HNC GTV segmentation has not really reached clinical practice yet. This might be due to lack of segmentation accuracy so far, in combination with the higher impact of a segmentation error on the dose distribution compared to OAR segmentation errors. In the first HECKTOR challenge on GTV segmentation (n = 201 for training), the winning team reached an average DSC of 0.76 and an $HD_{95\%}$ of 3.27 mm [2]. In the second challenge (n = 224 for training) the scores only slightly improved to DSC = 0.77 and $HD_{95\%}$=3.09 mm for the winning team [12]. It is important to note that these numbers are mean scores, which are not fully representative for performance on the entire test set. When results are presented in boxplots, there is a wide variety in scores, with a subset of patients with high DSC scores (>0.85), but also outliers towards much lower scores [13]. With the dataset of the 2nd HECKTOR challenge being only a small expansion of the first dataset, it will be interesting to follow what will happen in the 2022 challenge, where 524 patients are available for training.

The recent paper by Outeiral et al. [4] was based on an MR dataset of 230 oropharyngeal cancer patients (training n = 190, validation n = 20, testing n = 20), with each a 2D T1- and T2-weighted scan, and a 3D T1 weighted scan with gadolinium. The main objective was to address class imbalance challenges with two strategies. The first was to implement a fully automated 2-stage approach, where first a UNet was trained to localize a bounding box around the GTV, followed by training a UNet to segment the GTV using cropped data. The authors earlier published on the use of observer defined bounding boxes to improve segmentation results in a subset of this dataset [14]. The second approach was to compare four different loss functions, Dice-loss, Generalized-Dice-loss, Tversky-loss, and Unified-Focal-loss. These different loss functions were only evaluated in an end-to-end 3D UNet setting, without the image cropping. The main findings were that Generalized-Dice-Loss lead to the best results (median DSC = 0.54, $HD_{95\%}$ = 10.6 mm) but differences with other loss functions were not significant. The 2-stage approach further improved the segmentations to a median DSC = 0.64 and $HD_{95\%}$ = 8.7 mm.

The segmentation accuracy in this MR-only paper is not as good as what is produced in the HECKTOR challenge, where PET-CT data was used. Direct comparison is of course challenging, as these are different datasets of different patients and hospitals. Ren et al. recently showed in a multi-modal dataset of CT, PET, and MR data, that modality combinations which include the PET image (CT-PET, MR-PET, and CT-PET-MR) resulted in better DLS results compared to using CT-MR only [15]. This could of course be biased by which data is used to produce the ground truth delineation that is used for training and evaluation. If clinicians mainly look at the PET-CT data, and have the MRI on the side, the ground truth delineations might not fully cover the MR imaging information. However, in Outeiral et al. the ground truth GTVs were delineated on the 3D T1 gadolinium scans with the other MR modalities available, so using the same data that was used for DLS. In another recent paper, Wahid et al. investigated the use of different anatomical (T1, T2) and functional MR sequences (ADC, $K^{trans}$, $V_e$) on the segmentation accuracy [16]. On a small dataset of only 30 patients, they compared segmentation accuracy using a 3D residual UNet with T2 only, T2 plus T1, ADC, $K^{trans}$, or $V_e$, or all together in a leave-one-out cross-validation setting. Best results were obtained with T2 + T1, and interestingly, using all modalities together led to worse results than T2 only. It is of course questionable if a complex research question on different combinations of imaging modalities is sufficiently powered with only 30 patients. In a Turing test, three independent observers were not able to distinguish the DLS from the ground truth contour. More importantly, in a subjective assessment, 60 % of the DLS contours were deemed clinically acceptable, compared to only 64 % of the ground truth delineations. This finding illustrates the challenges in tumor segmentation in HNC in the first place, which is directly affecting the supervised DLS tools for tumors.

Regarding the 2-stage approach, it is interesting to compare the automated results with the previous paper using observer defined bounding boxes. In the first study, two observers manually placed a bounding box, with an average shift compared to the ground truth (bounding box of the ground truth tumor segmentation) of 3.0 and 8.9 mm [13]. This difference in bounding box accuracy led to a difference in segmentation accuracy of median DSC = 0.74 and $HD_{95\%}$ = 4.6 mm for observer 1 and median DSC = 0.67 and $HD_{95\%}$ = 7.2 mm. In the fully automated approach, the localization step had an average shift of 8.7 mm, resulting in a median DSC = 0.64 and $HD_{95\%}$ = 8.7 mm. The fully automated results are therefore more comparable to observer 2, and illustrate that the accuracy of the localization directly affects the accuracy of segmentation. It is also important to note that in the HECKTOR challenge, the imaging data is provided including an automatically generated bounding box around the oropharyngeal region.

In summary, DLS in HNC is a fast-developing research field, and expectations for the future are high. For OAR segmentation, DLS is used in more and more clinics, providing a great opportunity to assess performance in real-world-data. More research regarding the influence of dataset size is welcome, especially using more clinically relevant accuracy parameters such as $HD_{95\%}$ and MSD. The class imbalance problem is being addressed in many ways, and from the discussed literature, a combination of localization/cropping and loss-function optimization might be the way to go. For GTV segmentation, DLS tools are still mainly within the academic setting, and actual use in clinical practice is limited. The main challenge in this area is not in the dataset size or the class imbalance, it is in the quality of the ground truth segmentations.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## References

[1] Raudaschl PF, Zaffino P, Sharp GC, Spadea MF, Chen A, Dawant BM, et al. Evaluation of segmentation methods on head and neck CT: Auto-segmentation challenge 2015. Med Phys 2017;44:2020–36. https://doi.org/10.1002/mp.12197.
[2] Andrearczyk V, Oreiller V, Jreige M, Vallières M, Castelli J, Elhalawani H, et al. Overview of the HECKTOR challenge at MICCAI 2020: automatic head and neck tumor segmentation in PET/CT. Lect Notes Comput Sci (Including Subser Lect Notes Artif Intell Lect Notes Bioinformatics) 2021;12603 LNCS:1–21. https://doi.org/10.1007/978-3-030-67194-5_1.
[3] Henderson EGA, Vasquez Osorio EM, van Herk M, Green AF. Optimising a 3D convolutional neural network for head and neck computed tomography segmentation with limited training data. Phys Imaging Radiat Oncol 2022;22:44–50. https://doi.org/10.1016/j.phro.2022.04.003.
[4] Outeiral RR, Bos P, van der Hulst HJ, Al-Mamgani A, Jasperse B, Simões R, et al. Strategies for tackling the class imbalance problem of oropharyngeal primary tumor segmentation on magnetic resonance images. Phys Imaging Radiat Oncol 2022 (this volume).
[5] Nikolov S, Blackwell S, Zverovitch A, Mendes R, Livne M, De Fauw J, et al. Deep learning to achieve clinically applicable segmentation of head and neck anatomy

for radiotherapy. arXiv:1809.04430, 2021. https://doi.org/10.48550/arXiv.1809.04430.
[6] Tappeiner E, Welk M, Schubert R. Tackling the class imbalance problem of deep learning-based head and neck organ segmentation. Int J Comput Assist Radiol Surg 2022. https://doi.org/10.1007/s11548-022-02649-5.
[7] Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods 2021;18:203–11. https://doi.org/10.1038/S41592-020-01008-Z.
[8] Henderson E, Vasquez Osorio E, van Herk M, Brouwer C, Steenbakkers R, Green A. Accurate H&N 3D segmentation with limited training data using 2-stage CNNs (abstr). Radiother Oncol 2021;161:S1421–2. https://doi.org/10.1016/S0167-8140(21)08146-9.
[9] Fang Y, Wang J, Ou X, Ying H, Hu C, Zhang Z, et al. The impact of training sample size on deep learning-based organ auto-segmentation for head-and-neck patients. Phys Med Biol 2021;66:185012. https://doi.org/10.1088/1361-6560/ac2206.
[10] Brouwer CL, Boukerroui D, Oliveira J, Looney P, Steenbakkers RJHM, Langendijk JA, et al. Assessment of manual adjustment performed in clinical practice following deep learning contouring for head and neck organs at risk in radiotherapy. Phys Imaging Radiat Oncol 2020;16:54–60. https://doi.org/10.1016/j.phro.2020.10.001.
[11] Vaassen F, Boukerroui D, Looney P, Canters R, Verhoeven K, Peeters S, et al. Real-world analysis of manual editing of deep learning contouring in the thorax region. Phys Imaging Radiat Oncol 2022;22:104–10. https://doi.org/10.1016/j.phro.2022.04.008.
[12] Andrearczyk V, Oreiller V, Boughdad S, Rest CC Le, Elhalawani H, Jreige M, et al. Overview of the HECKTOR challenge at MICCAI 2021: automatic head and neck tumor segmentation and outcome prediction in PET/CT images. Lect Notes Comput Sci (Including Subser Lect Notes Artif Intell Lect Notes Bioinformatics) 2022;13209 LNCS:1–37. https://doi.org/10.1007/978-3-030-98253-9_1/TABLES/4.
[13] Oreiller V, Andrearczyk V, Jreige M, Boughdad S, Elhalawani H, Castelli J, et al. Head and neck tumor segmentation in PET/CT: The HECKTOR challenge. Med Image Anal 2022;77. https://doi.org/10.1016/J.MEDIA.2021.102336.
[14] Rodríguez Outeiral R, Bos P, Al-Mamgani A, Jasperse B, Simões R, van der Heide UA. Oropharyngeal primary tumor segmentation for radiotherapy planning on magnetic resonance imaging using deep learning. Phys Imaging Radiat Oncol 2021;19:39–44. https://doi.org/10.1016/j.phro.2021.06.005.
[15] Ren J, Eriksen JG, Nijkamp J, Korreman SS. Comparing different CT, PET and MRI multi-modality image combinations for deep learning-based head and neck tumor segmentation. Acta Oncol 2021;60:1399–406. https://doi.org/10.1080/0284186X.2021.1949034.
[16] Wahid KA, Ahmed S, He R, van Dijk LV, Teuwen J, McDonald BA, et al. Evaluation of deep learning-based multiparametric MRI oropharyngeal primary tumor auto-segmentation and investigation of input channel effects: Results from a prospective imaging registry. Clin Transl Radiat Oncol 2022;32:6–14. https://doi.org/10.1016/j.ctro.2021.10.003.

Jasper Nijkamp*
*Department of Clinical Medicine, Aarhus University, Aarhus, Denmark*
*Danish Center for Particle Therapy, Aarhus University Hospital, Aarhus, Denmark*

* Address: Department of Clinical Medicine, Aarhus University, Aarhus, Denmark.
*E-mail address:* jaspernijkamp@clin.au.dk.