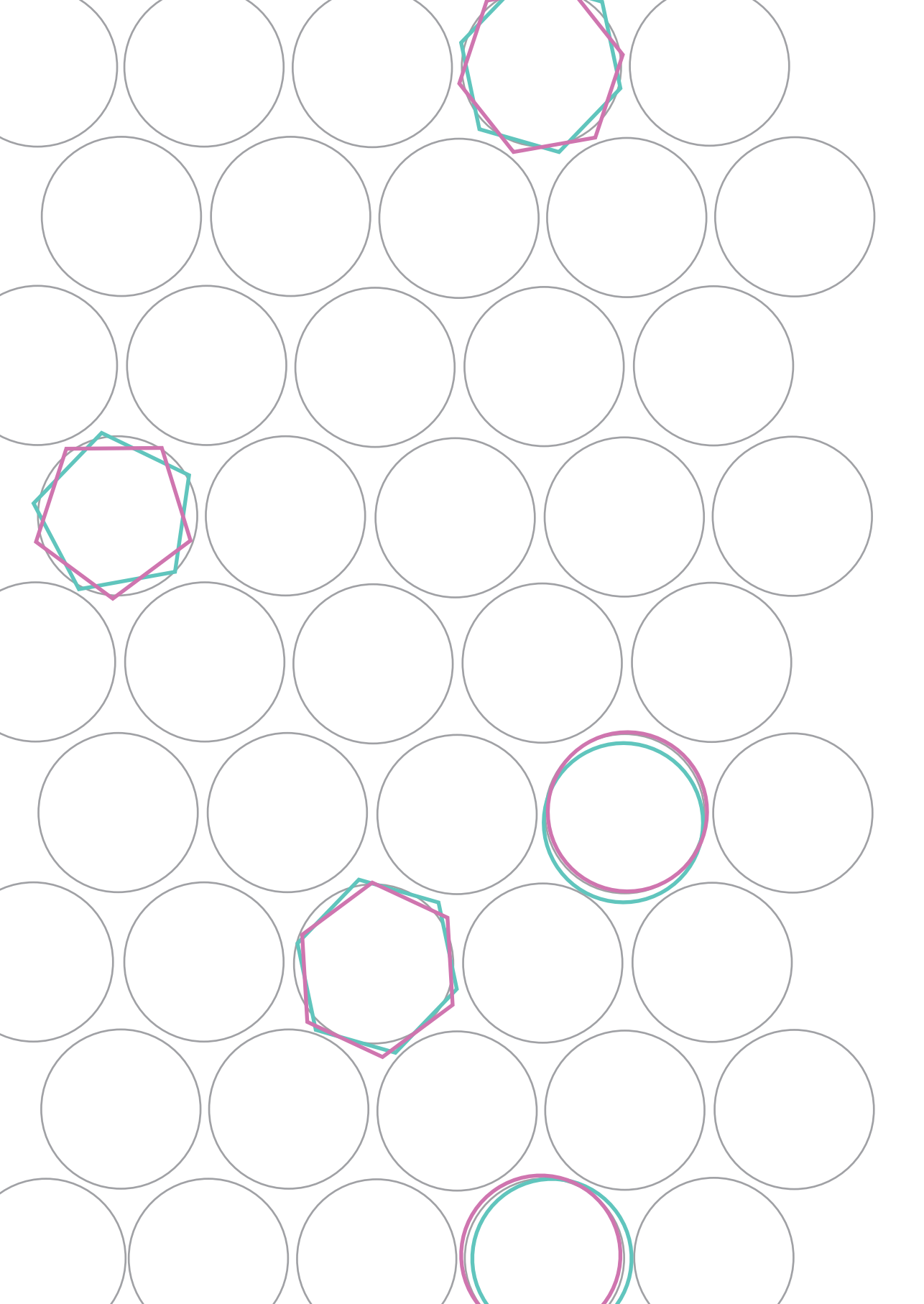
Deep learning for automatic segmentation of tumors on MRI

Rodríguez Outeiral, R.



Chapter 6

General discussion

In this thesis, we explored the topic of automated segmentation of tumors on MRI images. Deep learning (DL) techniques are already employed clinically in radiotherapy (RT) departments for organ-at-risk segmentation. However, tumor segmentation so far remains limited to a research setting. Furthermore, most existing studies about automatic segmentation of tumors use CT or FDG-PET images as the primary modality, even though MRI is often preferred to visualize cancer tissue in several tumor sites. We aimed to implement DL techniques to deliver clinically acceptable tumor segmentations in MRI cohorts. Two different MRI cohorts acquired in a clinical setting were used throughout this thesis: a cohort of oropharyngeal primary tumors in multiparametric diagnostic MRIs (in chapters 2 and 3) and a cohort of cervical cancer gross tumor volume in MRI images of brachytherapy treatment (in chapters 4 and 5).

INTRODUCING MULTIPLE MRI SEQUENCES AS INPUT FOR SEGMENTATION

When physicians manually segment tumors, they often rely on information from various sources, such as different imaging modalities or MRI sequences. Each of these images can provide distinct insights of the anatomy of the patient. The physicians can define the boundaries of the tumor by combining these different insights in their minds. Therefore, we also expect DL methods to benefit from combining different images as input.

In chapter 2, we investigated the effect on the segmentation performance of using different anatomical MRI sequences as input for the task of oropharyngeal cancer segmentation. The investigated MRI sequences were T1-weighted (T1w), T2-weighted (T2w) and T1 weighted after gadolinium injection (T1gd). We compared the segmentation performance of the networks trained with each of those sequences as input and with all the sequences together. Indeed, the network trained with all the available sequences outperformed the networks trained with one sequence only. This suggests that DL segmentation techniques also benefit from combining several input images.

Similar conclusions have been reached in other studies. Wahid et al. [42] investigated the use of both anatomical MRI sequences (T1w and T2w) and quantitative MRI sequences (ADC, Ktrans, and Ve) for automatic segmentation of the oropharyngeal tumor in an MRI-only workflow. Their study demonstrated that the best segmentations were achieved when combining both anatomical sequences. Ren et al. [80] explored the optimal combination of imaging modalities (MRI, CT, and PET) to improve the automatic segmentations. The model that incorporated the information from all the imaging modalities outperformed the rest of segmentation models. Thus, these works further support the claim that DL segmentation techniques benefit from being trained with multiple input images.

In all of the aforementioned studies, the quality of the automatic segmentations was assessed by comparing them to ground truth segmentations made by physicians. These ground truth segmentations were carried out on a single reference image. Although other images are consulted, the final voxel-level decision is based on just this reference image. This reference image may differ across different studies. For instance, in our work the ground truth segmentations were made in the T1gd sequence. In the study of Wahid et al [42], the ground truth segmentations were made in the T2w sequence. This can create differences in the ground truth segmentations. Given that the ground truth is used for training and evaluation of the network, it may also make the comparison between studies more challenging.

An alternative approach to compute the ground truth segmentations without the bias of a reference image is desirable. The presence of tumor at a pixel level could be confirmed with histopathology data, considered the gold standard in cancer diagnostics. However, obtaining this type of data is a complex process. It involves the surgical removal of the tumor tissue, subsequent staining, and precise registration with the 3D radiological images. Hence, histopathological validation is limited in the most tumor auto-segmentation studies, as acknowledged by Jager et al. [108].

In some tumor sites, the physicians also use the information from physical examinations to manually segment the tumor. This is the case for the two tumor sites studied throughout this thesis (i.e. the head and neck cancer and the cervical cancer), for which the tumor is reachable by the physicians. This information is not taken into account by the DL methods described in this section, which only rely on imaging data. The physicians in charge of the treatment might need to edit the automatic segmentation after the physical examination of the patient to include this information.

IMPROVING THE SEGMENTATIONS BY INCORPORATING PRIOR KNOWLEDGE

A common promise in the machine learning field is that underperforming methods would improve their performance by being trained on more data. As already stated in the introduction, inclusion of new data is challenging in the field of medicine. In the context of automatic segmentation of medical structures Fang et al. [109] observed that the performance improved logarithmically with the dataset size. They showed quantitatively that for the structures that were more dependent on the size of the dataset (i.e. the optical nerves, in their work) an improvement of 0.04 on the Dice Similarity Coefficient (DSC) was achieved with a training set 10 times larger. This suggests that even by collecting more data, performance gains may be rather modest. Therefore, a different approach to improve the segmentation performance is preferable.

In chapters 2 and 3, we hypothesized that the segmentation performance would improve by reducing the amount of context present in the image given as input. The rationale behind it is that by reducing the context around the tumor, the network does not need to spend any resources in localizing the tumor in the whole image. This simplifies the segmentation task, leaving more resources for the network to accurately segment the tumor. Our results confirmed that indeed, reducing the context led to improved performance.

Other strategies of simplifying the segmentation task have been investigated to improve the automatic segmentations in medical imaging. One approach is to use shape or anatomical constraints during training, thereby regularizing the solution space to only anatomically reasonable segmentations. This has been shown to result in more accurate segmentations than when training the segmentation network from scratch [110,111].

Another approach is to combine the segmentation task together with other relevant tasks, such as registration. In the context of adaptive image-guided radiotherapy for prostate cancer, S. Elmahdy et al. [112] framed the registration and segmentation as a joint problem within a multi-task learning setting. This yielded improved segmentation performance compared to the single-task setting.

Self-supervised learning (SSL) has been often posed as a promising strategy to improve the performance of machine learning techniques, particularly in low-data regimes [113,114]. SSL generally consists of leveraging information from unlabeled data during a pretraining phase. This approach reduces the need for extensive datasets for subsequent trainings. The work by Chaitanya et al [115]. serves as an example of the application of SSL to the field of automatic segmentation of medical images. In their study, they demonstrated improved segmentation performance for the networks trained with SSL compared to the networks initialized from scratch for three different medical segmentation tasks.

Leveraging the anatomical information of an individual patient by utilizing previous images and contours of that same patient can also improve the segmentations. This approach is sometimes referred to as "patient-specific fine-tuning" and is especially promising in scenarios that require segmentations for the same patient across different time points, such as the different fractions in RT. Li et al. [116] proposed such an approach for online contouring for MR-guided adaptive radiotherapy. Their results demonstrated improved segmentation performance compared to the network trained without patient-specific fine-tuning. Their segmentations were also more accurate than the segmentations generated by existing deformable registration algorithms commonly employed in clinical settings.

Though employing fairly different methodologies, these techniques share a common thread with the work presented in chapters 2 and 3: the integration of clinically relevant

prior knowledge during the training of segmentation networks is a promising strategy to improve the quality of the resulting automatic segmentations.

EVALUATION OF AUTOMATIC SEGMENTATIONS: WHAT IS A GOOD (AUTOMATIC) SEGMENTATION?

Determining the quality of the automatics segmentations was done in two ways throughout this thesis. In chapters 2 and 3, we geometrically compared the automatic segmentations to the manual segmentations made by expert radiologists. The aim of the radiologists was to accurately draw the extent of the tumor visible in the available imaging modalities. Therefore, a geometric comparison between the manual and automatic segmentations was adequate. This comparison was assessed with commonly used metrics: DSC, 95th Hausdorff distance (95th HD) and the mean surface distance (MSD).

In chapters 4 and 5, the manual segmentations used as ground truth for training and evaluation were performed by radiation oncologists. These segmentations were used in clinical practice to derive a treatment plan. A dosimetric evaluation was in this case also pertinent to evaluate the quality of the automatic segmentations. More specifically, we determined dose-volume parameters D90 and D98 for the automatic and expert delineations using the clinical dose distribution. Dosimetric evaluation of the automatic segmentation is important, because an error in the segmentation may have a different clinical impact depending on the dose that will be given to that point. This effect cannot be represented with geometric evaluation metrics, given that they do not take into account the dose distribution delivered to the patient.

In chapter 4, the segmentation performance of automatic segmentation of tumors in the cervix was further stratified in subgroups based on the tumor volume and the FIGO stage. Tumors with different FIGO stage or volume may appear differently on the MRI. Specifically, tumors with FIGO stage I will be limited to the cervix, while tumors with FIGO stage II and higher may extend to other anatomical structures, such as the vagina or the pelvic wall. The segmentation network potentially needs to look at different anatomical areas in the image to segment tumors depending on these clinical parameters. Therefore, differences in segmentation performance between the different subgroups of patients can arise. Another potential reason for performance differences is that some of these subgroups were under-represented in the training set. Regardless of the source of these differences, analyzing the segmentation performance separately for FIGO and volume can reveal a bias of the trained network towards certain subgroups of patients. Physicians could take this information into account to only use the segmentation network in the subgroups it performs best.

We used different evaluation metrics compared to other auto-segmentation works in literature. Firstly, it is not uncommon to find articles that report the DSC only. However, the DSC is volume-dependent by construction. Therefore, larger and more rounded structures typically result in higher values than smaller or more eccentric structures. This is particularly critical for structures with variable sizes, such as the tumors. Therefore we opted to always provide distance-based metrics together with the DSC. Secondly, the normalized surface distance (NSD) and the added path length (APL) are two quality metrics that are used in other works but not in this thesis. Although interesting metrics, they present shortcomings. The NSD is defined with a certain degree of tolerance, making them dependent on this hyper-parameter. The APL is not normalized or expressed with known units, making it less interpretable. Arguably, the distance-based metrics used in this work (95th HD, MSD and the surface DSC), also present shortcomings. The 95th HD primarily reflects the most significant error in the segmentation, often ignoring other relevant errors. MSD considers the entire contour but averages all errors together, which can lead to bias in the results, especially when gross errors are present. The surface DSC also depends on a tolerance parameter. To provide a more comprehensive assessment of contour quality, we recommend reporting a combination of these metrics.

Specifically for RT purposes, some geometric metrics may be more relevant than others. DSC has been shown not to correlate strongly with editing time [104] nor with dose/volume parameters [103]. In contrast, distance-based metrics, such as the 95th HD, the MSD, the surface DSC and the APL, have been shown to correlate more strongly with the editing time. Furthermore, a certain degree of geometric variability is expected in the manual tumor delineations due to interobserver variability. These geometric differences are taken into account by treatment margins for some tumor areas. If the error of the automatic segmentations is within the interobserver variability, the tumor segmentation might be clinically valid. Given that the treatment margins are defined as a certain distance around the tumor delineation, distance based metrics may also be more appropriate for this type of assessment.

Other works [42,117–119] have used the Turing test (or "Imitation game") to assess the quality of the automatic segmentations. The rationale behind it is that if a human observer cannot distinguish whether the contour was automatically generated or not, the contour closely resembles a manually created contour. Therefore, it is likely clinically acceptable. However, the source of the contour (automatic or human) does not necessarily indicate that the contour is clinically acceptable. Gooding et al. [117] proposed to complement the Turing test with additional questions that directly refer to clinical acceptability of the contour. These questions related to whether the observer would perform changes to the contour or how large those changes would be. In any case, these tests are strongly tied to

the expertise and environment of the observer. Consequently, we would recommend using them in complement with other evaluation techniques.

Overall, there are many approaches to evaluate the quality of a contour, each of them with their own advantages and limitations. Our recommendations would be two-fold: 1) to consider the clinical end goal of the contour, and use the evaluation framework that more closely assesses if that contour is acceptable, and 2) to combine the different evaluation metrics to describe the quality of the contour.

OPTIMIZING FOR THE RELEVANT LOSS FUNCTION

The loss function is a crucial part of the training of any neural network, including segmentation networks. Because of its importance, a multitude of different loss functions have been proposed in literature [30,70]. Ma et al. [30] grouped the available loss functions in three main categories: overlap-based loss functions, such as the DSC loss; distribution-based loss functions, such as the cross entropy loss or the focal loss; and boundary-based loss functions, such as the Hausdorff distance loss. Despite the large amount of loss functions available, the DSC loss is still selected in most works. A comprehensive evaluation of which loss function renders best segmentation performance for each specific task is rarely investigated.

In chapter 3 of this thesis, we trained the segmentation networks with different loss functions for the task of oropharyngeal cancer segmentation. The investigated loss functions were two overlap-based loss functions (DSC loss and Generalized DSC loss) and two distribution-based loss function (Focal Tversky loss and Unified Focal loss). No significant differences in DSC, 95th HD or MSD were found when training any of the loss functions. This suggests that the DSC loss was sufficient for our specific task. Similar conclusions were reached by Ma et al [30]. In their work, they compared 20 different loss functions on four different segmentation tasks. The showed that loss functions derived from the DSC were overall performing best.

All the loss functions compared in both our work and the work by Ma et al. reflect geometric differences between the automatic segmentation and the ground truth. This means that the segmentation networks are optimized during training to resemble the shape of the manual segmentations as closely as possible. However, the desired automatic segmentations for RT purposes are not necessarily those with the exact same shape than the manually acquired segmentations. In reality, the desired segmentations are those that reach the same dosimetric impact as the manual segmentations. Consequently, an interesting future line of research is to build loss functions that consider the dosimetric

impact in its definition. Further in the future, the loss function could even consider the RT patient outcomes in their definition, such as tumor control and toxicity.

QUALITY ASSURANCE OF AUTOMATIC SEGMENTATION OF TARGETS

Current automatic segmentation techniques do not render good quality segmentations for tumors in all the cases. This hinders their use in clinical practice, because physicians would still need to check, and potentially correct, the automatic segmentations. A possible solution is the implementation of quality assurance (QA) algorithms for the automatic segmentations. These QA algorithms could help the physicians by flagging the segmentations that would require review.

In chapter 5, we identified a metric that could be used for QA of the automatic segmentations. The proposed quality metric showed good capability to distinguish between cases that would require review as compared to adequate automatic segmentations that could be used without further check. Furthermore, the proposed metric correlated strongly with the DSC but also with clinically relevant distance-based metrics. These results indicate the potential of the metric as a QA tool.

Our work differed from the current literature of QA for automatic segmentations in two ways. Firstly, most works only correlate QA metrics to the DSC [32–34,102]. As discussed in previous sections, DSC does not fully describe the quality of the contour. Therefore, we considered it important to correlate our metric to distance-based metrics as well. Secondly, our metric can be derived directly from the output of the segmentation network. Instead, other works often extract their QA metrics from uncertainty maps [32,33,98], computed by applying the entropy operator on the output of the network. However, the entropy operator is not injective and can therefore destroy relevant information of the score maps.

Even if we can detect the segmentations that would require review, physicians would still need to spend time in the correction. A potential solution would consist on signaling the areas where the automatic segmentation is likely wrong. This information would assist the physicians in adjusting the contour in a semi-automatic manner. Even further in the future, the QA approaches could provide insights on how each potential editing would affect the final RT treatment outcome. With this relevant clinical information at hand, physicians could then make the final decision regarding the contour adjustments.

CLINICAL IMPLEMENTATION OF TUMOR AUTO-SEGMENTATION.

Ultimately, the aim is to implement automatic segmentation techniques for the tumors in clinical practice, as it is already the case for the OARs. As stated in the previous section, these techniques do not render acceptable segmentations in all the cases yet. Any errors in the tumor segmentation can have a large impact in the outcome of the patient, because tumors receive the highest radiation dose during the RT treatment. The responsibility of accepting the tumor segmentations lies with the radiation oncologists. However, they will not accept a segmentation unless they are certain of its quality.

The studies presented in this thesis can approach us towards clinical implementation in two ways. In chapters 2, 3 and 4, we focused on improving the quality of the automatic segmentations. By continuing to improve the auto-segmentations, we could eventually provide segmentations that radiation oncologists find acceptable in all cases. Furthermore, in chapter 5, we proposed a metric that can be used for QA purposes. With this tool in hand, the radiation oncologist could potentially distinguish which segmentations are clinically acceptable without the need of checking each auto-segmentation. However, further efforts are likely still needed to reach the clinical implementation of these techniques.

Vinod et al. [120] reviewed different applications to reduce the interobserver variability in target and OARs contouring. They observed that providing radiation oncologists with an automatic segmentation as a starting point that can be subsequently edited is an effective method to reduce contouring variability. One example of this was shown by Ferreira Silvério et al. [106] where they automatically segmented the mesorectum (CTV in MRI-guided rectal cancer treatment) and then asked an expert to manually correct the automatic segmentations. These corrections were not only comparable in terms of quality to the current clinical standard but also completed faster than the delineations made from scratch. Therefore, the clinical implementation of automatic segmentation techniques could already provide clinical value as an aid to the radiation oncologist to edit in a semi-automatic setting. Arguably, such an evaluation framework is a necessary step towards the clinical implementation of auto-segmentation tools. This approach can provide clinically acceptable segmentations more quickly than the current clinical approaches, while also increasing the trust in the auto-segmentations.

CONCLUSIONS

In this thesis, we explored the topic of automatic segmentation of tumors on MRI. Deep learning methods are already employed in clinical settings to segment anatomical structures. However, automatic segmentation of tumors remains in a research capacity, showcasing the complexity of the problem. A promising approach to improve the quality of the segmentations consists of utilizing prior information to guide the training of the segmentation network. Two examples have been demonstrated in this thesis: the reduction of context around the tumor and the incorporation of different MRI sequences. Furthermore, clinical relevant information should be considered both during the training and evaluation of these methods. During training, this can be achieved by defining clinically relevant loss functions. During the evaluation, this is possible by defining clear clinical end points. Besides potential improvements in the quality of the automatic segmentations, automatic QA can also play a crucial role in advancing towards clinical applicability. QA methods are not only important for safety reasons but also to increase the trust of clinicians in these techniques. Finally, a currently viable strategy involves an interactive approach in which a candidate auto-segmentation is provided as a starting point. This could presently reduce the time spent by the clinical staff in the segmentations, thereby enhancing the efficiency of the RT workflow.