Dose Guidance for Radiotherapy-Oriented Deep Learning Segmentation

Elias Rüfenacht[1(✉)] , Robert Poel[2], Amith Kamath[1] , Ekin Ermis[2],
Stefan Scheib[3], Michael K. Fix[2], and Mauricio Reyes[1,2]

[1] ARTORG Center for Biomedical Engineering Research,
University of Bern, Bern, Switzerland
`elias.ruefenacht@unibe.ch`
[2] Department of Radiation Oncology, Inselspital, Bern University Hospital
and University of Bern, Bern, Switzerland
[3] Varian Medical Systems Imaging Laboratory GmbH, Baden, Switzerland

Abstract. Deep learning-based image segmentation for radiotherapy is intended to speed up the planning process and yield consistent results. However, most of these segmentation methods solely rely on distribution and geometry-associated training objectives without considering tumor control and the sparing of healthy tissues. To incorporate dosimetric effects into segmentation models, we propose a new training loss function that extends current state-of-the-art segmentation model training via a dose-based guidance method. We hypothesized that adding such a dose-guidance mechanism improves the robustness of the segmentation with respect to the dose (i.e., resolves distant outliers and focuses on locations of high dose/dose gradient). We demonstrate the effectiveness of the proposed method on Gross Tumor Volume segmentation for glioblastoma treatment. The obtained dosimetry-based results show reduced dose errors relative to the ground truth dose map using the proposed dosimetry-segmentation guidance, outperforming state-of-the-art distribution and geometry-based segmentation losses.

Keywords: Segmentation · Radiotherapy · Dose Guidance · Deep Learning

1 Introduction

Radiotherapy (RT) has proven effective and efficient in treating cancer patients. However, its application depends on treatment planning involving target lesion and radiosensitive organs-at-risk (OAR) segmentation. This is performed to guide radiation to the target and to spare OAR from inappropriate irradiation. Hence, this manual segmentation step is very time-consuming and must

be performed accurately and, more importantly, must be patient-safe. Studies have shown that the manual segmentation task accounts for over 40% of the treatment planning duration [7] and, in addition, it is also error-prone due to expert-dependent variations [2,24]. Hence, deep learning-based (DL) segmentation is essential for reducing time-to-treatment, yielding more consistent results, and ensuring resource-efficient clinical workflows.

Nowadays, training of DL segmentation models is predominantly based on loss functions defined by geometry-based (e.g., SoftDice loss [15]), distribution-based objectives (e.g., cross-entropy), or a combination thereof [13]. The general strategy has been to design loss functions that match their evaluation counter-part. Nonetheless, recent studies have reported general pitfalls of these metrics [4,19] as well as a low correlation with end-clinical objectives [11,18,22,23]. Furthermore, from a robustness point of view, models trained with these loss functions have been shown to be more prone to generalization issues. Specifically, the Dice loss, allegedly the most popular segmentation loss function, has been shown to have a tendency to yield overconfident trained models and lack robustness in out-of-distribution scenarios [5,14]. These studies have also reported results favoring distribution-matching losses, such as the cross-entropy being a strictly proper scoring rule [6], providing better-calibrated predictions and uncertainty estimates. In the field of RT planning for brain tumor patients, the recent study of [17] shows that current DL-based segmentation algorithms for target structures carry a significant chance of producing false positive outliers, which can have a considerable negative effect on applied radiation dose, and ultimately, they may impact treatment effectiveness. In RT planning, the final objective is to produce the best possible radiation plan that jointly targets the lesion and spares healthy tissues and OARs. Therefore, we postulate that training DL-based segmentation models for RT planning should consider this clinical objective.

In this paper, we propose an end-to-end training loss function for DL-based segmentation models that considers dosimetric effects as a clinically-driven learning objective. Our contributions are: (i) a dosimetry-aware training loss function for DL segmentation models, which (ii) yields improved model robustness, and (iii) leads to improved and safer dosimetry maps. We present results on a clinical dataset comprising fifty post-operative glioblastoma (GBM) patients. In addition, we report results comparing the proposed loss function, called **Do**se-**Se**gmentation **Lo**ss (DOSELO), with models trained with a combination of binary cross-entropy (BCE) and SoftDice loss functions.

2   Methodology

Figure 1 describes the general idea of the proposed DOSELO. A segmentation model (U-Net [20]) is trained to output target segmentation predictions for the Gross Tumor Volume (GTV) based on patient MRI sequences. Predicted segmentations and their corresponding ground-truth (GT) are fed into a dose predictor model, which outputs corresponding dose predictions (denoted as $\widehat{D}_P$ and D_P in Fig. 1). A pixel-wise mean squared error between both dose predictions is then
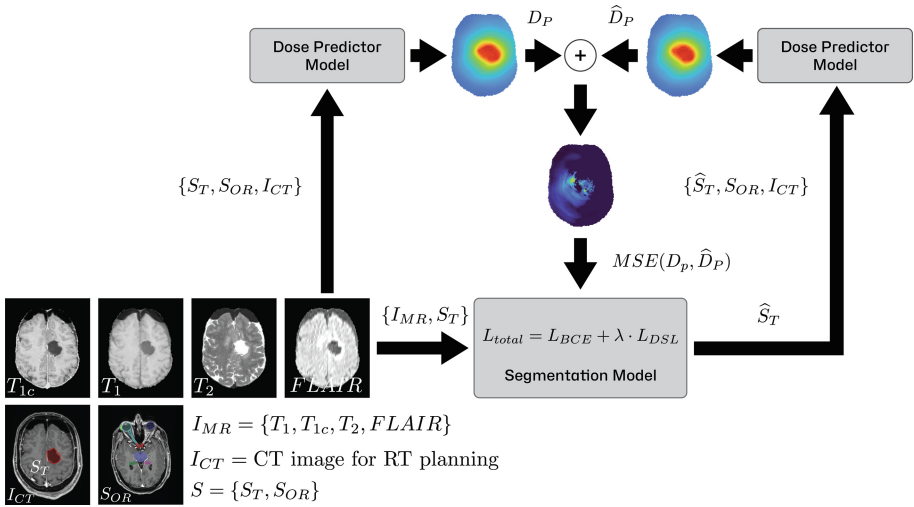
Fig. 1. Schematic overview of the proposed dosimetry-aware training loss function. A segmentation model (U-Net [20]) is trained to output target segmentation predictions ($\widehat{S}_T$) for the Gross Tumor Volume (GTV) based on patient MRI sequences I_{MR}. Predicted ($\widehat{S}_T$) and ground-truth segmentations (S_T) are fed into the dose predictor model along with the CT-image (I_{CT}), and OAR segmentation (S_{OR}). The dose predictor outputs corresponding dose predictions $\widehat{D}_P$ and D_P. A pixel-wise mean squared error between both dose predictions is calculated, and combined with the binary cross-entropy (BCE) loss to form the final loss, $L_{total} = L_{BCE} + \lambda L_{DSL}$.

calculated and combined with the BCE loss to form the final loss. In the next sections we describe the adopted dose prediction model [9,12], and the proposed DOSELO.

2.1   Deep Learning-Based Dose Prediction

Recent DL methods based on cascaded U-Nets have demonstrated the feasibility of generating accurate dose distribution predictions from segmentation masks, approximating analytical dose maps generated by RT treatment planning systems [12]. Originally proposed for head and neck cancer [12], this approach has been recently extended for brain tumor patients [9] with levels of prediction error below 2.5 Gy, which is less than 5% of the prescribed dose. This good level of performance, along with its ability to yield near-instant dose predictions, enables us to create a training pipeline that guides learned features to be dose-aware.

Following [12], the dose predictor model consists of a cascaded U-Net (i.e., the input to the second U-Net is the output of the first concatenated with the input to the first U-Net) trained on segmentation masks, CT images, and reference dose maps. The model's input is a normalized CT volume and segmentation masks for target volume and OARs. As output, it predicts a continuous-valued dose map of the same dimension as the input. The model is trained via deep

supervision as a linear combination of L2-losses from the outputs of each U-Net in the cascade. We refer the reader to [9,12] for further implementation details. We remark that the dose predictor model was also trained with data augmentation, so imperfect segmentation masks and corresponding dose plans are included. This allows us in this study to use the dose predictor to model the interplay between segmentation variability and dosimetric changes.

Formally, the dose prediction model M_D receives as inputs: segmentations masks for the GTV $S_T \in \mathbb{Z}^{W \times H}$ and the OARs $S_{OR} \in \mathbb{Z}^{W \times H}$, the CT image (used for tissue attenuation calculation purposes in RT) $I_{CT} \in \mathbb{R}^{W \times H}$, and outputs $M_D(S_T, S_{OR}, I_{CT}) \mapsto D_P \in \mathbb{R}^{W \times H}$, a predicted dose map where each pixel value in D corresponds to the local predicted dose in Gy. Due to the limited data availability, we present results using 2D-based models but remark that their extension to 3D is straightforward. Working in 2D is also feasible from an RT point of view because the dose predictor is based on co-planar volumetric modulated arc therapy (VMAT) planning, commonly used in this clinical scenario.

2.2   Dose Segmentation Loss (DOSELO)

During the training of the segmentation model, we used the dose predictor model to generate pairs of dose predictions for the model-generated segmentations and the GT segmentations. The difference between these two predicted dose maps is used to guide the segmentation model. The intuition behind this is to guide the segmentation model to yield segmentation results being dosimetrically consistent with the dose maps generated via the corresponding GT segmentations.

Formally, given a set of N pairs of labeled training images $\{(I_{MR}, S_P)_i : 1 \leq i \leq N\}$, $I_{MR} \in \mathbb{R}^D$ (with $D : \{T1, T1c, T2, FLAIR\}$ MRI clinical sequences), and corresponding GT segmentations of the GTV $S_T \in \mathbb{Z}^{H \times W}$, a DL segmentation model $M_S(I_{MR}) \mapsto \widehat{S}_T$ is commonly updated by minimizing a standard loss term, such as the BCE loss (L_{BCE}).

To guide the training process with dosimetry information stemming from segmentation variations, we propose to use the mean squared error (MSE) between dose predictions for the GT segmentation (S_T) and the predicted segmentation ($\widehat{S}_T$), and construct the following dose-segmentation loss,

$$L_{DSL} = \frac{1}{H \times W} \sum_i^{H \times W} (D_P^i - \widehat{D}_P^i)^2 \tag{1}$$

$$D_P = M_D(S_T, S_{OR}, I_{CT}) \tag{2}$$

$$\widehat{D}_P = M_D(\widehat{S}_T, S_{OR}, I_{CT}), \tag{3}$$

where D_P^i and $\widehat{D}_P^i$ denote pixel-wise dose predictions. The final loss is then,

$$L_{total} = L_{BCE} + \lambda L_{DSL}, \tag{4}$$

where λ is a hyperparameter to weigh the contributions of each loss term. We remark that during training we use standard data augmentations including spatial transformations, which are also subjected to dose predictions, so the model is informed about relevant segmentation variations producing dosimetry changes.

3 Experiments and Results

3.1 Data and Model Training

We divide the descriptions of the two separate datasets used for the dose prediction and segmentation models.

Dose Prediction: The dose prediction model was trained on an in-house dataset comprising a total of 50 subjects diagnosed with post-operative GBM. This includes CT imaging data, segmentation masks of 13 OARs, and the GTV. GTVs were defined according to the ESTRO-ACROP guidelines [16]. The OARs were contoured by one radiotherapist according to [21] and verified by mutual consensus of three experienced radiation oncology experts. Each subject had a reference dose map, calculated using a standardized clinical protocol with Eclipse (Varian Medical Systems Inc., Palo Alto, USA). This reference was generated on basis of a double arc co-planar VMAT plan to deliver 30 times 2 Gy while maximally sparing OARs. We divided the dataset into training (35 cases), validation (5 cases), and testing (10 cases). We refer the reader to [9] for further details.

Segmentation Models: To develop and test the proposed approach, we employed a separate in-house dataset (i.e., different cases than those used to train the dose predictor model) of 50 cases from post-operative GMB patients receiving standard RT treatment. We divided the dataset into training (35 cases), validation (5 cases), and testing (10 cases). All cases comprise a planning CT registered to the standard MRI images (T1-post-contrast (Gd), T1-weighted, T2-weighted, FLAIR), and GT segmentations containing OARs as well as the GTV. We note that for this first study, we decided to keep the dose prediction model fixed during the training of the segmentation model for a simpler presentation of the concept and modular pipeline. Hence, only the parameters of the segmentation model are updated.

Baselines and Implementation Details: We employed the same U-Net [20] architecture for all trained segmentation models, with the same training parameters but two different loss functions, to allow for a fair comparison. As a strong comparison baseline, we used a combo-loss formed by BCE plus SoftDice, which is also used by nnUNet and recommended by its authors [8]. This combo-loss has also been reported as an effective one [13]. For each loss function, we computed a

five-fold cross-validation. Our method[1] was implemented in PyTorch 1.13 using Adam optimizer [10] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, batch normalization, dropout set at 0.2, learning rate set at 10^{-4}, $2 \cdot 10^4$ update iterations, and a batch size of 16. The architecture and trained parameters were kept constant across compared models. Training and testing were performed on an NVIDIA Titan X GPU with 12 GB RAM. The input image size is 256×256 pixels with an isotropic spacing of 1 mm.

3.2   Evaluation

To evaluate the proposed DOSELO, we computed dose maps for each test case using a standardized clinical protocol with Eclipse (Varian Medical Systems Inc., Palo Alto, USA). We calculated dose maps for segmentations using the state-of-the-art BCE+SoftDice and the proposed DOSELO. For each obtained dose map, we computed the dose score [12], which is the mean absolute error between the reference dose map (D_{S_T}) and the dose map derived from the corresponding segmentation result ($D_{\widehat{S}_T}$, where $\widehat{S}_T \in \{$BCE+SoftDice, DOSELO$\}$), and set it relative to the reference dose map (D_{S_T}) (see Eq. 5).

$$RMAE = \frac{1}{H \times W} \sum_i^{H \times W} \frac{|D_{S_T} - D_{\widehat{S}_T}|}{D_{S_T}} \tag{5}$$

Although it has been shown that geometric-based segmentation metrics poorly correlate with the clinical end-goal in RT [4,11,18,23], we report in supplementary material Dice and Hausdorff summary statistics as well (supplementary Table 3). We nonetheless reemphasize our objective to move away from such proxy metrics for RT purposes and promote the use of more clinically-relevant ones.

3.3   Results

Figure 2 shows results on the test set, sorted by their dosimetric impact. We found an overall reduction of the relative mean absolute error (RMAE) with respect to the reference dose maps, from 0.449 ± 0.545, obtained via the BCE+SoftDice combo-loss, to 0.258 ± 0.201 for the proposed DOSELO (i.e., an effective 42.5% reduction with $\lambda = 1$). This significant dose error reduction shows the ability of the proposed approach to yield segmentation results in better agreement with dose maps obtained using GT segmentations than those obtained using the state-of-the-art BCE+SoftDice combo-loss.

Table 1 shows results for the first and most significant four cases from a RT point of view (due to space limitations, all other cases are shown in supplementary material). We observe the ability of the proposed approach to significantly reduce outliers, generating a negative dosimetry impact on the dose
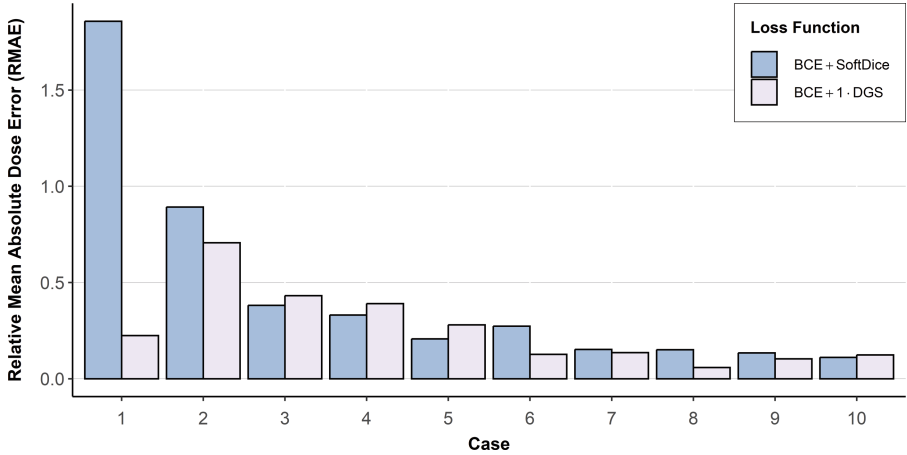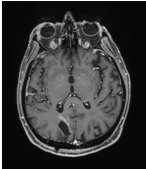
---

[1] Code available under https://github.com/ruefene/doselo.
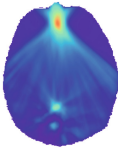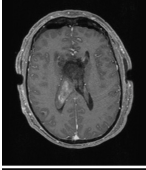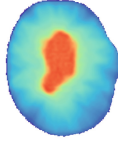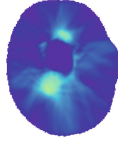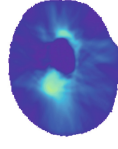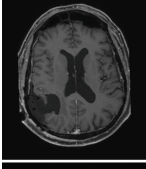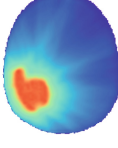
Fig. 2. Relative mean absolute dose errors/differences (RMAE) between the reference dose map and dose maps obtained using the predicted segmentations. Lower is better. Acro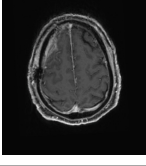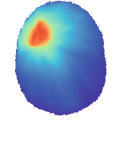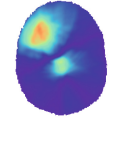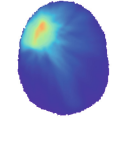ss all tested cases and folds we observe a large RMAE reduction for dose maps using the proposed DOSELO (average RMAE reduction of 42.5%).

maps. We analyzed case number 3, 4, and 5 from Fig. 2 for which the standard BCE+SoftDice was slightly better than the proposed DOSELO. For case no. 3 the tumor presents a non-convex shape alongside the skull's parietal lobe, which was not adequately modeled by the training dataset used to train the segmentation models. Indeed, we remark that both models failed to yield acceptable segmentation quality in this area. In case no. 4, both models failed to segment the diffuse tumor area alongside the skull; however, as shown in Fig. 2-case no. 4, the standard BCE+SoftDice model would yield a centrally located radiation dose, with strong negative clinical impact to the patient. Case no. 5 (shown in supplementary material) is an interesting case called butterfly GBM, which is a rare type of GBM (around 2% of all GBM cases [3]), characterized by bihemispheric involvement and invasion of the corpus callosum. In this case, the training data also lacked characterization for such cases. Despite this limitation, we observed favorable dose distributions with the proposed method.

Although we are aware that classical segmentation metrics poorly correlate with dosimetric effects [18], we report that the proposed method is more robust than the baseline BCE+SoftDice loss function, which yields outliers with Hausdorff distances: 64.06 ± 29.84 mm vs 28.68 ± 22.25 mm (-55.2% reduction) for the proposed approach. As pointed out by [17], segmentation outliers can have a detrimental effect on RT planning. We also remark that the range of HD values is in range with values reported by models trained using much more training data (see [1]), alluding to the possibility that the problem of robustness might not be directly solvable with more data. Dice coefficients did not deviate significantly between the baseline and the DOSELO models (DSC: 0.713 ± 0.203 (baseline) vs. 0.697 ± 0.216 (DOSELO)).

Table 1. Comparison of dose maps and their absolute differences to the reference dose maps (BCE+SoftDice (BCE+SD), and the proposed DOSELO). It can be seen that DOSELO yields improved dose maps, which are in better agreement with the reference dose maps (dose map color scale: 0 (blue) - 70Gy (red)).

4 Discussion and Conclusion

The ultimate goal of DL-based segmentation for RT planning is to provide reliable and patient-safe segmentations for dosimetric planning and optimally targeting tumor lesions and sparing of healthy tissues. However, current loss functions used to train models for RT purposes rely solely on geometric considerations that have been shown to correlate poorly with dosimetric objectives [11,18,22,23]. In this paper, we propose a novel dosimetry-aware training loss function, called DOSELO, to effectively guide the training of segmentation models toward dosimetric-compliant segmentation results for RT purposes. The proposed DOSELO uses a fast-dose map prediction model, enabling model guidance on how dosimetry is affected by segmentation variations. We merge this information into a simple yet effective loss function that can be combined with existing ones. These first results on a dataset of post-operative GBM patients show the

ability of the proposed DOSELO to deliver improved dosimetric-compliant segmentation results. Future work includes extending our database of GBM cases and to other anatomies, as well as verifying potential improvements when co-training the segmentation and dose predictor models, and jointly segmenting GTVs and OARs. With this study, we hope to promote more research toward clinically-relevant DL training loss functions.

References

1. Bakas, S., et al.: Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. arXiv preprint arXiv:1811.02629 (2018)
2. Cloak, K., et al.: Contour variation is a primary source of error when delivering post prostatectomy radiotherapy: results of the trans-tasman radiation oncology group 08.03 radiotherapy adjuvant versus early salvage (raves) benchmarking exercise. J. Med. Imag. Radiat. Oncol. **63**(3), 390–398 (2019)
3. Dayani, F., et al.: Safety and outcomes of resection of butterfly glioblastoma. Neurosurg. Focus **44**(6), E4 (2018)
4. Fidon, L., et al.: A dempster-shafer approach to trustworthy AI with application to fetal brain MRI segmentation. arXiv preprint arXiv:2204.02779 (2022)
5. Galdran, A., Carneiro, G., Ballester, M.A.G.: On the optimal combination of cross-entropy and soft dice losses for lesion segmentation with out-of-distribution robustness. In: Yap, M.H., Kendrick, C., Cassidy, B. (eds.) Diabetic Foot Ulcers Grand Challenge. DFUC 2022. LNCS, vol. 13797. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-26354-5_4
6. Gneiting, T., Raftery, A.E.: Strictly proper scoring rules, prediction, and estimation. J. Am. Stat. Assoc. **102**(477), 359–378 (2007)
7. Guo, C., Huang, P., Li, Y., Dai, J.: Accurate method for evaluating the duration of the entire radiotherapy process. J. Appl. Clin. Med. Phys. **21**(9), 252–258 (2020)
8. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods **18**(2), 203–211 (2021)
9. Kamath, A., Poel, R., Willmann, J., Andratschke, N., Reyes, M.: How sensitive are deep learning based radiotherapy dose prediction models to variability in organs at risk segmentation? In: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI), pp. 1–4. IEEE (2023)
10. Kingma, D., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
11. Kofler, F., et al.: Are we using appropriate segmentation metrics? Identifying correlates of human expert perception for CNN training beyond rolling the dice coefficient. arXiv preprint arXiv:2103.06205 (2021)
12. Liu, S., Zhang, J., Li, T., Yan, H., Liu, J.: A cascade 3D U-Net for dose prediction in radiotherapy. Med. Phys. **48**(9), 5574–5582 (2021)
13. Ma, J., et al.: Loss odyssey in medical image segmentation. Med. Image Anal. **71**, 102035 (2021)
14. Mehrtash, A., Wells, W.M., Tempany, C.M., Abolmaesumi, P., Kapur, T.: Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. IEEE Trans. Med. Imaging **39**(12), 3868–3878 (2020). https://doi.org/10.1109/TMI.2020.3006437

15. Milletari, F., Navab, N., Ahmadi, S.A.: V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D vision (3DV), pp. 565–571. IEEE (2016)
16. Niyazi, M., et al.: ESTRO-ACROP guideline "target delineation of glioblastomas." Radiotherapy Oncol. **118**(1), 35–42 (2016)
17. Poel, R., et al.: Impact of random outliers in auto-segmented targets on radiotherapy treatment plans for glioblastoma. Radiat. Oncol. **17**(1), 170 (2022)
18. Poel, R., et al.: The predictive value of segmentation metrics on dosimetry in organs at risk of the brain. Med. Image Anal. **73**, 102161 (2021)
19. Reinke, A., et al.: Common limitations of performance metrics in biomedical image analysis. In: Medical Imaging with Deep Learning (2021)
20. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
21. Scoccianti, S., et al.: Organs at risk in the brain and their dose-constraints in adults and in children: a radiation oncologist's guide for delineation in everyday practice. Radiother. Oncol. **114**(2), 230–238 (2015)
22. Vaassen, F., et al.: Evaluation of measures for assessing time-saving of automatic organ-at-risk segmentation in radiotherapy. Phys. Imag. Radiat. Oncol. **13**, 1–6 (2020)
23. Vandewinckele, L., et al.: Overview of artificial intelligence-based applications in radiotherapy: recommendations for implementation and quality assurance. Radiother. Oncol. **153**, 55–66 (2020)
24. Vinod, S.K., Jameson, M.G., Min, M., Holloway, L.C.: Uncertainties in volume delineation in radiation oncology: a systematic review and recommendations for future studies. Radiother. Oncol. **121**(2), 169–179 (2016)