Contents lists available at ScienceDirect

Physics and Imaging in Radiation Oncology

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

Original Research Article

Towards interactive deep-learning for tumour segmentation in head and neck cancer radiotherapy

Zixiang Wei [a,b], Jintao Ren [a,b], Stine Sofia Korreman [a,b,c], Jasper Nijkamp [a,b,*]

[a] Aarhus University, Department of Clinical Medicine, Aarhus, Denmark
[b] Danish Center for Particle Therapy, Aarhus University Hospital, Aarhus, Denmark
[c] Department of Oncology, Aarhus University Hospital, Aarhus, Denmark

ARTICLE INFO

Keywords:
Head and neck cancer
Radiotherapy
Tumour segmentation
Interactive deep-learning
Convolutional neural network

ABSTRACT

Background and purpose: With deep-learning, gross tumour volume (GTV) auto-segmentation has substantially been improved, but still substantial manual corrections are needed. With interactive deep-learning (iDL), manual corrections can be used to update a deep-learning tool while delineating, minimising the input to achieve acceptable segmentations. We present an iDL tool for GTV segmentation that took annotated slices as input and simulated its performance on a head and neck cancer (HNC) dataset.
Materials and methods: Multimodal image data of 204 HNC patients with clinical tumour and lymph node GTV delineations were used. A baseline convolutional neural network (CNN) was trained ($n = 107$ training, $n = 22$ validation) and tested ($n = 24$). Subsequently, user input was simulated on initial test set by replacing one or more of predicted slices with ground truth delineation, followed by re-training the CNN. The objective was to optimise re-training parameters and simulate slice selection scenarios while limiting annotations to maximally-five slices. The remaining 51 patients were used as an independent test set, where Dice similarity coefficient (DSC), mean surface distance (MSD), and 95% Hausdorff distance ($HD_{95\%}$) were assessed at baseline and after every update.
Results: Median segmentation accuracy at baseline was DSC = 0.65, MSD = 4.3 mm, $HD_{95\%} = 17.5$ mm. Updating CNN using three slices equally sampled from the craniocaudal axis of the GTV in the first round, followed by two rounds of annotating one extra slice, gave the best results. The accuracy improved to DSC = 0.82, MSD = 1.6 mm, $HD_{95\%} = 4.8$ mm. Every CNN update took 30 s.
Conclusions: The presented iDL tool achieved substantial segmentation improvement with only five annotated slices.

1. Introduction

In radiotherapy, target volume delineation is still suffering from substantial inter-observer variation (IOV) [1–3]. To improve delineation accuracy and reduce workload, researchers have focused on deep-learning (DL) based auto-segmentation [4]. While DL tools dominate the leaderboards in medical image segmentation challenges, the accuracies for tumour segmentation are not yet perfect. For head and neck cancer (HNC) gross tumour volume (GTV) segmentation, uni- and multi-modal imaging inputs have been investigated, such as computerized tomography (CT)-only [5], magnetic resonance (MR)-only [6,7], positron emission tomography (PET)-CT [8–10], and PET-CT-MR [11], with segmentation accuracies ranging from 0.6 to 0.8 in Dice similarity coefficient (DSC). In clinical practice, it is expected that these segmentation tools would still require substantial manual annotations from physicians. Most DL tools are based on supervised learning, where imaging data with manual segmentation is used to train a convolutional neural network (CNN). With the presence of IOV in the training data, segmentation accuracy is limited by the level of the IOV [12].

Instead of fully automated segmentation, some researchers proposed interactive deep-learning (iDL) methods combining the power of CNNs with the knowledge of the physicians. In its most simple form, iDL can be a manually placed bounding box around the volume of interest to focus the inference step [7,13–15]. In more advanced iDL, the physician is

* Corresponding author at: Palle Juul Jensensboulevard 25, 8200 Aarhus, Denmark.
E-mail addresses: zixiang-wei@clin.au.dk (Z. Wei), jintaoren@oncology.au.dk (J. Ren), stine.korreman@oncology.au.dk (S.S. Korreman), jaspernijkamp@clin.au.dk (J. Nijkamp).

https://doi.org/10.1016/j.phro.2022.12.005
Received 4 August 2022; Received in revised form 19 December 2022; Accepted 22 December 2022
Available online 24 December 2022

presented with an initial auto-segmentation and clicks, scribbles, or drag points are used to indicate false positive (FP) or false negative (FN) areas. These are then used to retrain the CNN parameters, to improve the segmentation output for the case at hand. Scribble-based iDL has been successfully demonstrated in pancreas segmentation on CT [16], organs at risk and tumour segmentation on MRI [17], and also as a tool to annotate data in the training process [18]. While interactions are easy and fast to set, physicians are always first presented with an auto-segmentation for annotation, which could bias the delineation process. An approach where the physician provides a limited amount of surface points [19] or a few slice contours to guide the auto-segmentation might prevent this segmentation bias.

In this study, we developed an iDL tool which learns from physicians while they delineate the target volume, improving the segmentation accuracy on the go, minimizing the need for user input to achieve acceptable delineations. At first, one or a few contoured slices provided by the physician are used to update the CNN parameters to become more patient and observer specific. Subsequently, an improved segmentation of the entire tumour is presented, followed by repeated interactive contour annotation and CNN retraining if needed.

2. Materials and methods

We first trained a baseline CNN, randomly splitting our dataset into training (n = 107), validation (n = 22), and test (n = 24) sets. Thereafter, we simulated user interaction on the initial test set by replacing a predicted tumour contour on selected slices with the ground truth contour. The simulations were used to optimise the hyperparameters for the iDL tool and to systematically assess how the selection of slices affected the segmentation accuracy. Finally, the optimised hyperparameters were used to simulate interactive segmentation on an independent test set (n = 51). For each simulated iDL patient, we started with the baseline CNN, meaning that the iDL was only used to optimise the CNN for the patient at hand and reset afterwards.

In this study, we focused on the improvement of segmentation accuracy with each learning iteration and how many user inputs were needed. We did not optimise the time needed for re-training. We aimed to achieve segmentations which could be clinically acceptable within five annotated slices.

2.1. Dataset

Multimodal imaging data of 204 HNC patients treated with (chemo-) radiotherapy between 2013 and 2020 was used [11]. Data from 2015 to 2018 (n = 153) were used for training, validation, and testing. Remaining data (n = 51) was used as the independent test set. Use of data for this study was approved by The Danish National Committee on Health Research Ethics (Reference number: 2018311), informed consent was waived. All patients had a planning FDG-PET/CT scan, an axial T2 weighted MR scan and a coronal T1 weighted in-phase MDIXON, all in an immobilisation mask in treatment position. The two MR scans were deformably registered to the CT scan using Elastix [20], and all data was sliced on the CT grid. The union of the primary tumour volume (GTVt) and the involved nodes volume (GTVn) was used as the ground truth. Details on the dataset can be found in Supplementary Material 1.

2.2. CNN architecture and loss functions

UNet is a widely used segmentation CNN with the advantage of providing high precision with relatively low data amount demand [21,22]. UNet++ is a variant of UNet, which can be seen as an efficient ensemble of UNets of varying depths, which allows you to avoid extensive architecture depth searches [23,24]. In this study, we compare UNet++ and UNet to investigate their abilities in the context of iDL. With our goal in mind to interactively train the CNN based on a small set of provided or annotated contours on slices, 2D CNNs were used, providing a straightforward architecture to handle the input data (further details in Supplementary Material 2).

Loss functions for DL segmentation often encompass one of the segmentation accuracy measures, such as DSC in Dice-loss. As the target region only encompasses a small portion of the image leading to class imbalance, weighted Dice-loss is commonly used [25]. However, due to variations in target size between patients, Dice-loss might be unstable and cause severe oscillation in the loss curve during the training [11]. Besides, when using a 2D CNN, it is also important to handle the data imbalance, as many slices will only have a background label. Focal-loss [26], a variant of Cross-entropy-loss, can deal with class imbalance and data imbalance with more stable output on the loss curve during training. Finally, Dice-loss can be combined with Focal-loss to get the best of both worlds, for example, in Hybrid Focal-loss [27]. In this study, we compared Dice-loss, Focal-loss, and Hybrid Focal-loss.

2.3. Baseline training

The four modalities (CT, PET, MR-T1, and MR-T2) were concatenated as CNN input data in four channels. Both UNet and UNet++ architectures were trained using the three different loss functions and Adam optimiser. An overview of the optimised hyperparameters can be found in Supplementary Table 3.1. The training was set for 100 epochs, with a batch size of 100, and early stopping when the validation loss did not improve for 30 epochs. The validation set was used for fine-tuning. The final evaluations using different CNN architectures and loss functions were performed on the initial test set, reporting DSC, mean surface distance (MSD), and 95% Hausdorff distance ($HD_{95\%}$).

2.4. Interactive deep-learning

iDL simulations were performed on the initial test set (n = 24) to optimise the iDL hyperparameters. In every simulation, we started with a baseline prediction. Subsequently, an interaction was simulated by replacing the predicted contour with the ground truth on one or more slices, followed by a fine-tuned re-training of the CNN using the annotated slices only. An update of the CNN is termed “a round” in the remainder of the paper.

2.4.1. Slice selection scenarios

The amount of provided annotated data in each round might influence the performance [16]. Therefore, we simulated different slice selection scenarios where in the first round, the update of the CNN was based on N = 1, 2, 3, 4, or 5 slices. In the subsequent rounds, we always added one slice at a time and simulated up to 10 annotated slices in total. The selection was limited to slices that contained either a prediction, a ground truth segmentation, or both. The following scenarios were simulated:

(1) “Largest”: select *N*-slices with the largest predicted tumour area in the first round, followed by adding one slice at a time (largest non-annotated prediction) in the next round.

(2) “Equal-divide”: find *N*-slices on the cranial-caudal axis that divided the available selection into n + 1 equal parts in the first round, followed by the largest non-annotated slice in subsequent rounds.

(3) “Random”: randomly select *N*-slices in the first round, followed by random selection of one non-annotated slice in subsequent rounds.

2.4.2. Data augmentation

In each round, the CNN parameters were updated based on limited simulated input data. To prevent overfitting, we used the same data augmentation methods as baseline training (Supplementary Table 4.3.1) and further optimised how many augmentations (augmentation-times = 1, 2, 4, 8, and 16) of each input slice were needed (Supplementary Table 4.3.2).

From the second round, the annotated slices of former round(s) were also used for re-training the CNN to prevent knowledge loss. To balance

Table 1
Mean segmentation accuracy using different baseline dropout-rates and CNN architectures.

CNN	Baseline dropout-rate	DSC				MSD (mm)				$HD_{95\%}$ (mm)			
		Baseline	Round 1	Round 2	Round 3	Baseline	Round 1	Round 2	Round 3	Baseline	Round 1	Round 2	Round 3
UNet	0.00	0.70	0.79	0.81	0.82	**5.6**	3.5	3.4	3.0	**30.7**	17.1	13.6	13.3
	0.40	**0.71**	**0.80**	**0.82**	**0.83**	6.0	**1.9**	**1.4**	**1.3**	36.4	**6.7**	**5.2**	**4.5**
UNet++	0.00	0.69	0.80	0.82	0.83	7.1	2.4	1.6	1.5	39.8	9.4	4.8	4.4
	0.30	**0.70**	**0.81**	**0.83**	**0.84**	**5.3**	**1.5**	**1.3**	**1.2**	**25.1**	**5.6**	**4.6**	**4.3**

The results in Table 1 were obtained on the initial test set. Here, the "Equal-divide" scenario with N = 3,1,1 was used. Bold values indicate the best score per baseline/round per parameter.

the influence of slices of former rounds with the current round, we used an augmentation decay strategy, where augmentation-times of a previous round were divided by two.

2.4.3. Assessment of results

First, the optimised iDL hyperparameters and slice selection scenario was determined on the initial test set, using DSC, MSD, $HD_{95\%}$ of each round up to the moment where five annotated slices were used for retraining. Subsequently, the actual performance was evaluated on the

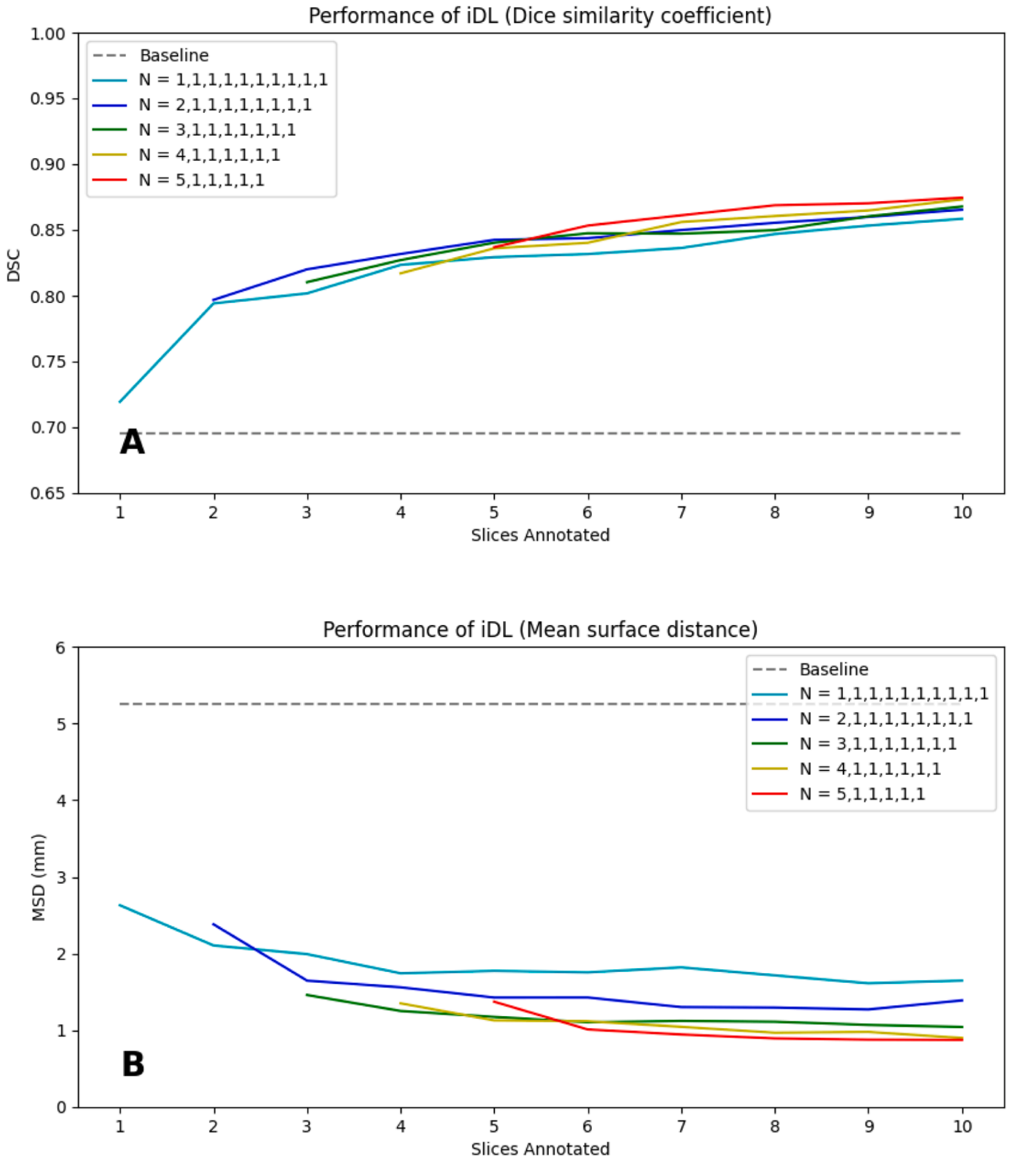

Fig. 1. Mean segmentation accuracy on the initial test set in terms of (A) DSC and (B) MSD when using a different number of slices in the first round of iDL. In the subsequent rounds, one slice was annotated at a time, up to ten in total. The performance of the baseline network is depicted with the dotted line. The results of $HD_{95\%}$ can be found in Supplementary Fig. 4.1.

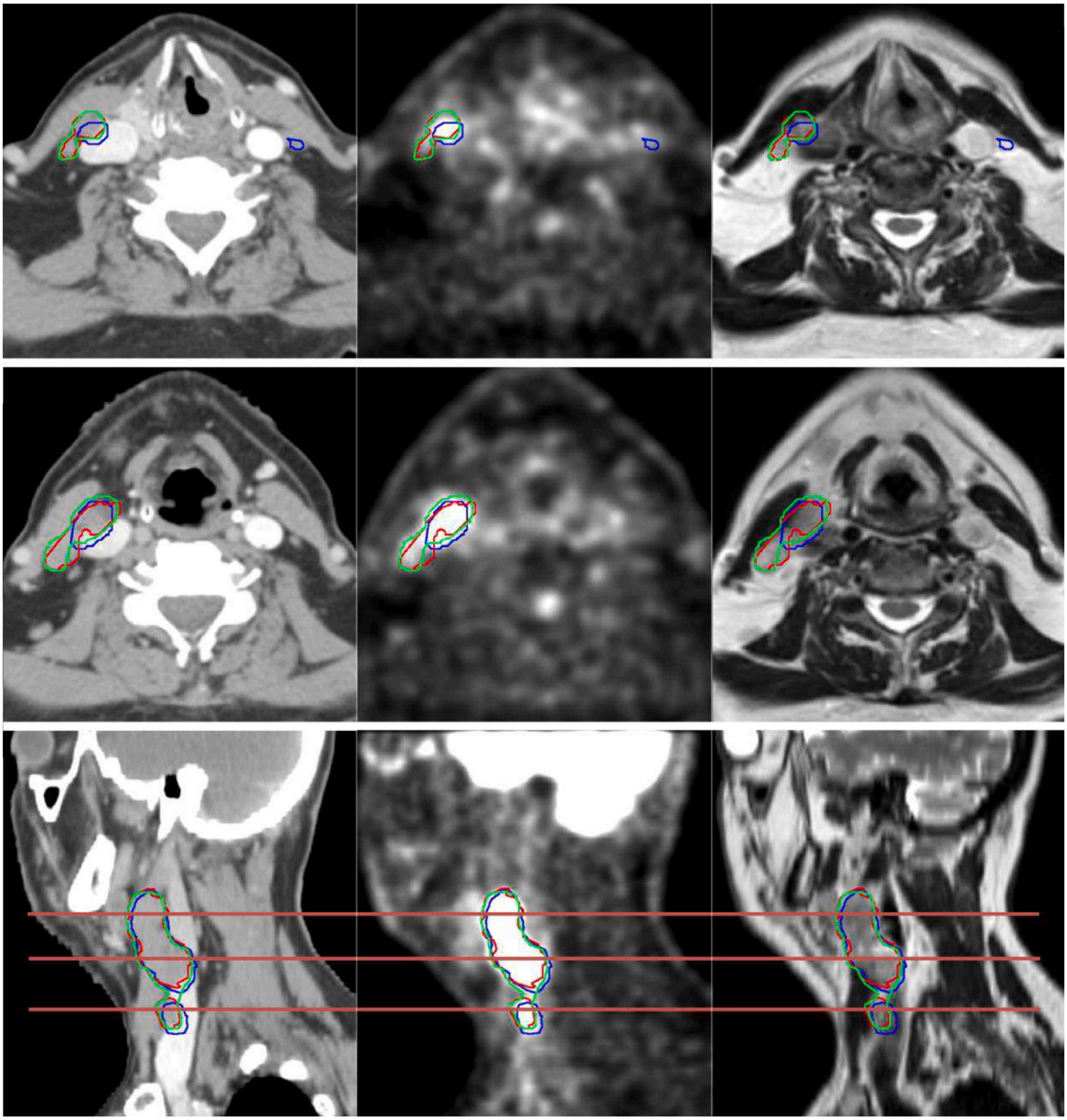

Fig. 2. Example case of one round of iDL updating using three slices in the "Equal-divide" scenario. Each row shows a slice of the CT (left), PET (middle) and T2 weighted MRI (right). The ground truth is shown in red, baseline segmentation in blue, and the iDL updated segmentation in green. The top row shows an annotated axial slice. The middle row shows a non-annotated axial slice. The bottom row shows a sagittal slice, where the orange horizontal lines indicate the axial slices that were updated in the iDL process. The axial images in the top row correspond to the most caudal updated slice. The axial images in the middle row were from the slice in the middle between the most caudal updated slice and the central updated slice. Baseline segmentation metrics were: DSC = 0.78, MSD = 3.9 mm, $HD_{95\%}$ = 11.5 mm. After one round of iDL, the segmentation metrics improved to: DSC = 0.86, MSD = 0.8 mm, $HD_{95\%}$ = 2.6 mm. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

independent test set (n = 51) using the same hyperparameters. In this work, we assumed that retraining based on the annotated slices improved the overall tumour segmentation accuracy, also on the slices that were not annotated. To test this assumption, we calculated the DSC at baseline and every update round while excluding the five slices that were annotated in the update rounds. These DSC values were compared between baseline and each round using a Wilcoxon Signed-Rank Test. To put our work into perspective, we have also trained the nnUNet [28] using our dataset.

3. Results

3.1. Baseline training

Details on how the different CNN architectures, loss functions and hyperparameters influenced the segmentation accuracy and the final selected baseline hyperparameters can be found in Supplementary Table 3.1. For baseline training, results using UNet were marginally better than with UNet++ (Supplementary Table 3.2 and baseline results in Table 1). The best results in terms of DSC were obtained with Hybrid Focal-loss, while Focal-loss provided the best MSD and $HD_{95\%}$ results. Dropout-rates of 0.3–0.4 provided a small additional improvement in

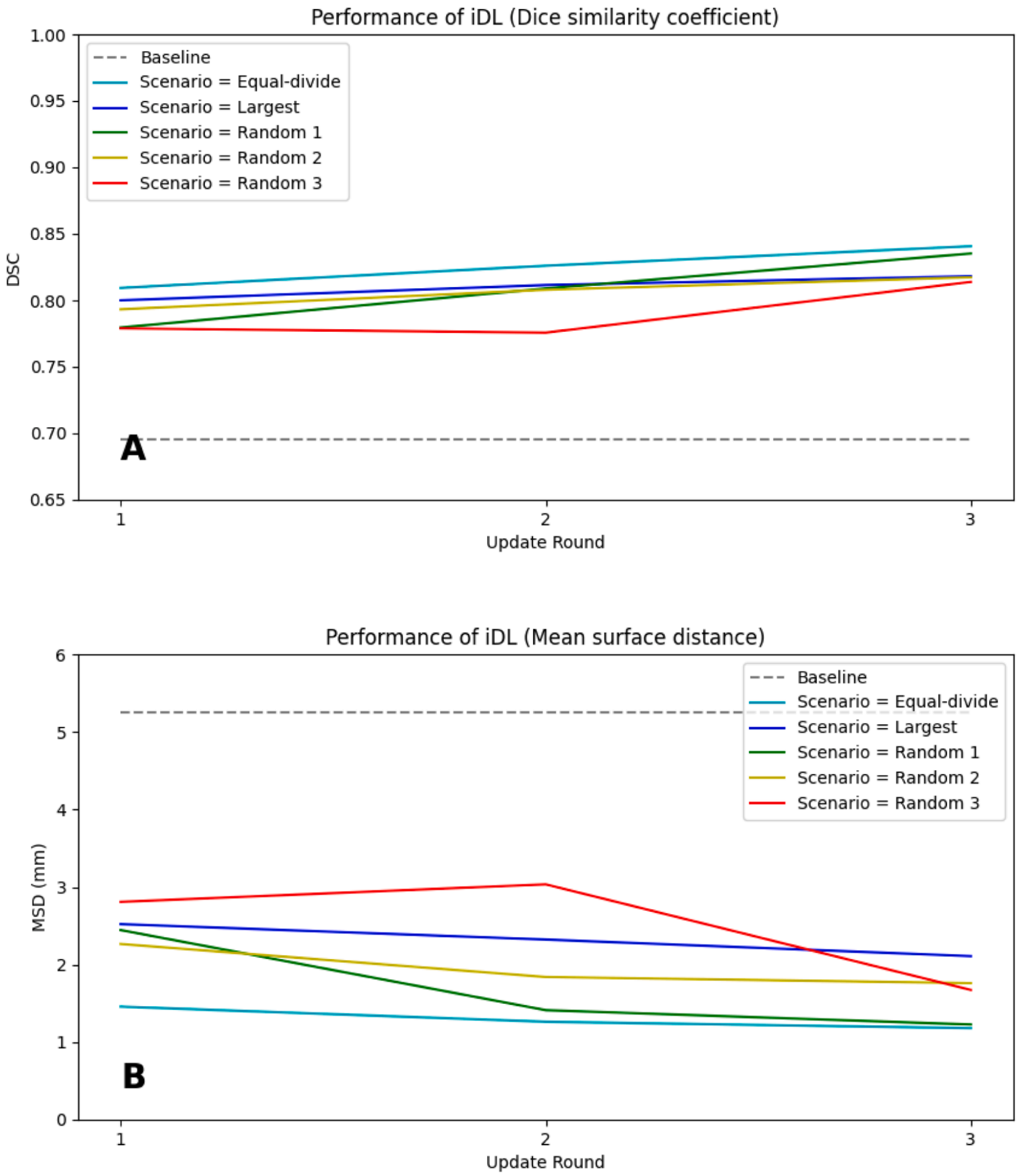

Fig. 3. Mean segmentation accuracy on the initial test set in terms of (A) DSC and (B) MSD when using different scenarios in selecting the slices to update for iDL. In this simulation, three slices were selected in round one, followed by adding one slice at a time in rounds two and three. The performance of the baseline CNN is depicted with the dotted line. $HD_{95\%}$ results can be found in Supplementary Fig. 4.2.

segmentation accuracy (Supplementary Table 3.3).

3.2. Simulation of iDL

3.2.1. Effect of architecture and baseline dropout-rate on iDL

We selected Hybrid Focal-loss for the remainder of the study since it achieved the best DSC results. Given the small differences between UNet and UNet++, both architectures were evaluated for iDL. To illustrate how baseline dropout-rate affects iDL, we simulated iDL using the best slice selection scenario: “Equal-divide” (see below) and using three annotated slices in the first round for both the UNet and UNet++ architecture with and without using dropout (Table 1). Regardless of the CNN architecture, iDL using a baseline trained with dropout obtained better results than using a non-dropout baseline. Furthermore, UNet++ showed improved performance in the iDL phase over UNet in terms of MSD and $HD_{95\%}$. Therefore, UNet++ with a dropout-rate of 0.3 was used in the remainder of the study.

3.2.2. Annotating N-slices in the first round

In the five simulations where the annotation in the first round was based on N = 1, 2, 3, 4, or 5 slices, and then one slice at a time up to ten in total, we ran all simulations using “Equal-divide” as the scenario with fixed hyperparameters. For annotation of up to five slices in total, scenarios that annotated three or four slices in the first round obtained the best segmentation results, especially when combining DSC, MSD, and $HD_{95\%}$ (Fig. 1 and Supplementary Fig. 4.1). Beyond five slices, the DSC increased a little further. However, the decline in MSD and $HD_{95\%}$ was limited, and for some scenarios, numbers actually increased again beyond eight slices annotated. Therefore, in the remainder of the study simulations, we always used three slices in the first round, as it balances effort and effect for the physician in the first step. An example of how the segmentation improved after round one using three annotated slices is

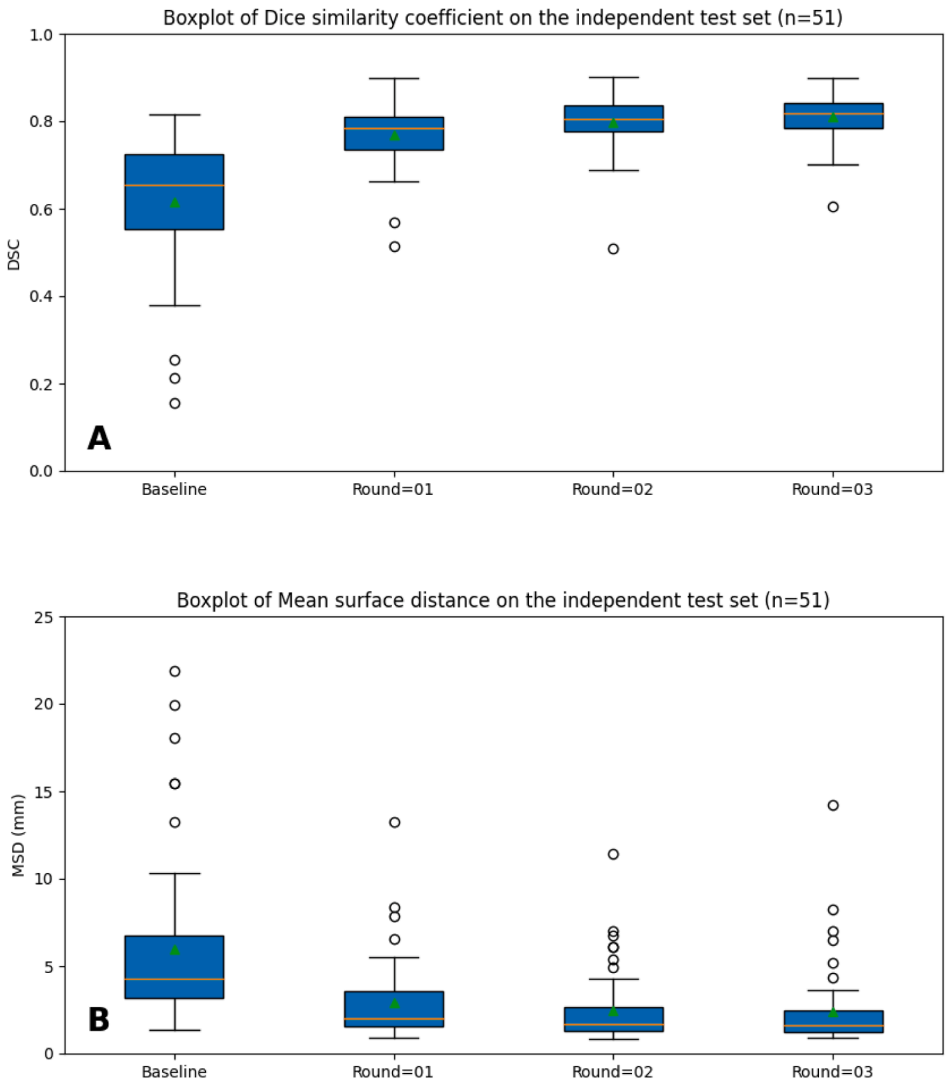

Fig. 4. Boxplots of the segmentation accuracy on the independent test set in terms of (A) DSC and (B) MSD when using the optimised iDL hyperparameters. The boxplots present the median (horizontal line), the mean (triangle), the central 50% (solid box), and the range of data excluding outliers (whiskers). $HD_{95\%}$ results can be found in Supplementary Fig. 4.4.1.

shown in Fig. 2.

3.2.3. *Comparison of different selection scenarios*

Of the slice selection scenarios, “Equal-divide” was the best choice in all accuracy metrics (Fig. 3). From the three simulations using the “Random” scenario, it was obvious that it mattered which slices were chosen. In the “Largest” scenario, often adjacent slices were selected as the input, resulting in an underrepresentation of the smaller parts of the GTV.

3.2.4. *iDL performance on the independent test set*

With the optimised iDL hyperparameters (Supplementary Table 4.3.1), we annotated five slices over three rounds (N = 3,1,1), and iDL took 35 s, 32 s, and 28 s in rounds one, two, and three, respectively. The independent test set median baseline results were DSC = 0.65, MSD = 4.3 mm, $HD_{95\%}$ = 17.5 mm, improving to DSC = 0.82, MSD = 1.6 mm, $HD_{95\%}$ = 4.8 mm after three update rounds (Fig. 4). The median DSC score for baseline and the three rounds excluding the five annotated slices were 0.65, 0.75, 0.76, and 0.77, respectively (each improvement was statistically significant: $p < 0.0001$, $p < 0.0001$, and $p = 0.0008$, respectively). Two example cases from the independent test set where the iDL tool did not perform as expected are shown in Fig. 5. The segmentation results using nnUNet are shown in Supplementary Material 5.

4. Discussion

We presented a novel iDL tumour segmentation tool that took individual annotated slices to improve segmentation performance during a delineation session. In our simulations, the UNet++ architecture trained with Hybrid Focal-loss and a dropout-rate at baseline training of 0.3 provided the best results. It was also clear that there was a strong dependence on which and how many slices were annotated in the iDL process. With only five slices annotated using an equal-divide strategy in three rounds, the median segmentation results improved from DSC = 0.65, MSD = 4.3 mm, $HD_{95\%}$ = 17.5 mm at baseline to DSC = 0.82, MSD = 1.6 mm, $HD_{95\%}$ = 4.8 mm.

The closest available study on HNC tumour iDL segmentation is by Outeiral et al., who tested the use of a region of interest around the

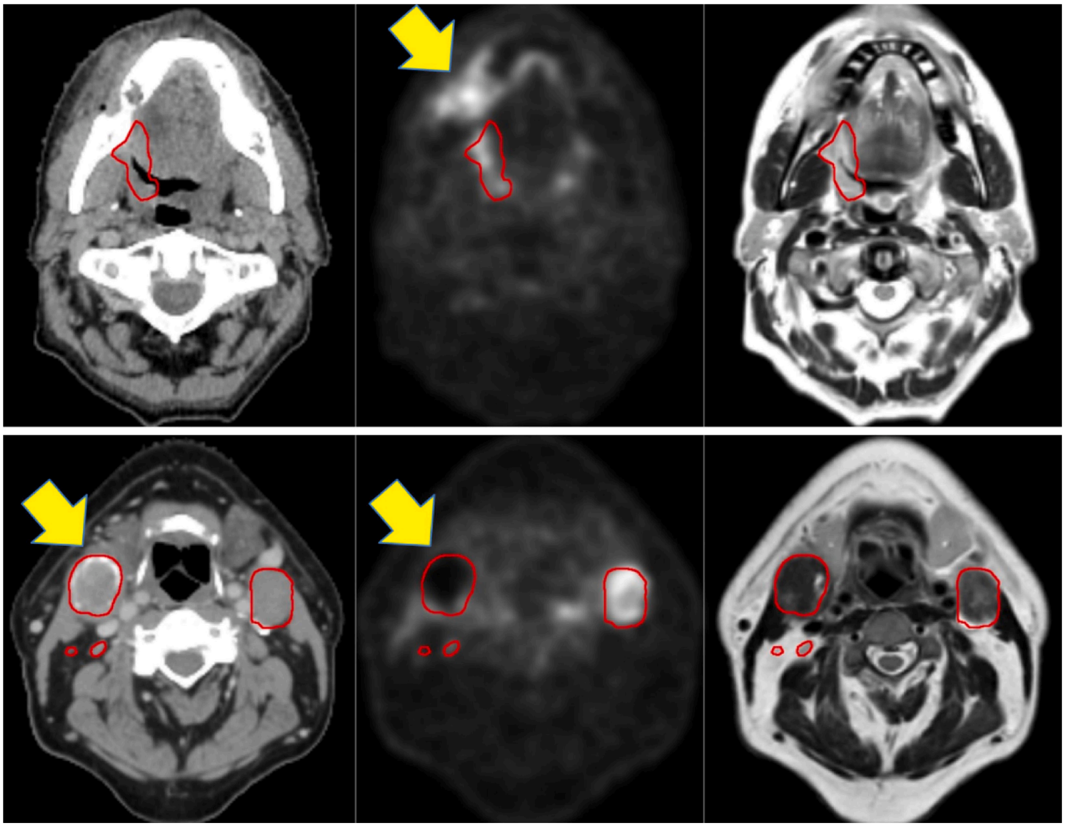

Fig. 5. Example cases from the independent test set which were challenging for the iDL tool. Each row shows the planning CT (left), PET scan (middle), T2 MR (right), and ground truth delineation (Red). The top row shows a patient case with a jaw defect that caused inflammation, resulting in a PET-SUV of 9 (yellow arrow in the PET image (middle). Baseline prediction for this case was DSC = 0.25, MSD = 8.0 mm, $HD_{95\%}$ = 20.6 mm. This slice was included in the first iDL update round, showing a moderate increase of the DSC score to 0.57, but $HD_{95\%}$ increased to 24.5 mm, as parts of the brain were included in the GTV. The bottom row shows a case with bilateral lymph node involvement, where the large node on the right side of the patient was necrotic (no uptake on the PET, see yellow arrow), including a calcified ring on the CT scan. This slice was updated in round three and resulted in accurate segmentation of both the PET positive nodes and the PET negative node, but also in many false positive regions in the patient's lower neck. This mainly affected the MSD and $HD_{95\%}$, which were 3.2 mm and 14.6 mm after round two, increasing to 14.2 mm and 88.3 mm, respectively, after round three. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

tumour area to improve tumour segmentation on MR scans. At baseline, they reached a median DSC of 0.55, MSD of 2.7 mm, $HD_{95\%}$ of 8.7 mm, which improved to DSC = 0.74, MSD = 1.2 mm, $HD_{95\%}$ = 4.6 mm, and DSC = 0.67, MSD = 1.7 mm, $HD_{95\%}$ = 7.2 mm for two observers, respectively. Direct comparison with this study is challenging, as baseline results and imaging modalities are different. Still, it shows at least that limited interaction with a clinician can substantially improve segmentation accuracy. Furthermore, both studies show that the improvement is dependent on the input of the observers. The closest paper regarding iDL technique is the study by Boers et al. [16], who used scribble-based updates in pancreas segmentation on CT imaging. In their study, the DSC improved with the same approximate magnitude, and the total segmentation time was reduced by a factor of 2 compared to manual segmentation. It is unfortunately not clear how many scribbles were needed to reach an acceptable segmentation or if multiple scribble rounds were needed to get the wanted contour. One advantage of the iDL tool presented here over scribble-based interaction is that it can be used in a blinded fashion for the first round of annotation, preventing some potential segmentation bias.

When assessing the segmentation accuracy in our study, the baseline results were worse than using the nnUNet (Supplementary Material 5). The difference is most likely due to the use of a singlefold 2D-UNet in the current study versus a 5-fold ensemble 3D-UNet. However, with only three slices annotated, the results were already at par with nnUNet in terms of DSC and $HD_{95\%}$, improving to significantly better results from iDL round 2. This indicates that with iDL, it is possible to annotate the predicted segmentation to the patient at hand quickly. Similar to our previous work [11], the PET scan was essential to the segmentation accuracy, especially at baseline. This is illustrated in Supplementary Fig. 4.4.2, where the baseline segmentation accuracy was worse for patients with a low GTV tumour SUV_{mean} (below SUV 4). After one update round, most cases improved substantially to a DSC between 0.65 and 0.9. From Fig. 5, it is also clear that rare imaging cases can have a major impact on the iDL segmentation accuracy. Several steps can be made to further improve the baseline prediction. For example, using an ensemble of segmentations from k random initialized models, as well as test-time augmentation for more robust segmentations to get rid of FP and FN [29].

We have clearly demonstrated that the improvement in segmentation using iDL was dependent on which and how many slices were annotated for each round (Fig. 1 and Fig. 3). When updating based on only one slice, results didn't really change, which can be caused by the overfitting of the network to the sparse data provided, as shown earlier by Boers et al. [16]. Some of the scenarios presented in Fig. 1 resulted in a reduction in accuracy after five slices. Since we selected the largest area slices after the first round, this might have resulted in a selection of adjacent slices, making the segmentation locally better, but losing accuracy further away at smaller areas of the GTV.

In this study, UNet and UNet++ performed similarly in baseline training, but UNet++ performed best in the interactive part (Table 1), especially in MSD and $HD_{95\%}$. Potentially, the redesigned skip connections, which aggregate features of varying semantic scales of UNet++, attributed to the improved performance. The thought behind these extra connections was that feature maps from the decoder and encoder networks become more semantically similar, which simplifies the learning task for the optimiser [23,24].

This study came with several limitations. First of all, interactions were simulated in this study, mainly to create a setup in which architecture and hyperparameter choices could be optimised systematically. User tests with realistic behaviour will be needed to assess how the iDL tool performs in clinical practice. It might be needed to supply specific user guidelines on how the tool performs best. To make the iDL tool applicable in the clinic, we also need to address the speed of optimisation. One round now took approximately 30 s, which might challenge the patience of the physicians. One obvious speedup could come from introducing weight maps to focus the iDL optimisation on FP and FN areas, reducing the number of needed iterations [16].

In conclusion, we have presented a slice-based iDL segmentation tool with the intention of improving auto-segmentation accuracy with limited input from observers. In the simulations, annotating three to five slices in one to three rounds substantially improved the segmentation accuracy.

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.

Acknowledgements

We would like to thank chief consultants Jesper Grau Eriksen and Kenneth Jensen for their keen insights into manual GTV segmentation from the oncologist perspective.

Funding

This work was supported by:
Danish Cancer Society [grant no. R231-A13856]
PhD school of Health, Aarhus University, Denmark
DCCC Radiotherapy – The Danish National Research Center for Radiotherapy, Danish Cancer Society [grant no. R191-A11526]
Novo Nordisk Foundation [grant no. NNF195A0059372]

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.phro.2022.12.005.

References

[1] Cardenas CE, Blinde SE, Mohamed ASR, Ng SP, Raaijmakers C, Philippens M, et al. Comprehensive quantitative evaluation of variability in magnetic resonance-guided delineation of oropharyngeal gross tumor volumes and high-risk clinical target volumes: an R-IDEAL stage 0 prospective study. Int J Radiat Oncol 2022;113:426–36. https://doi.org/10.1016/j.ijrobp.2022.01.050.

[2] Njeh CF. Tumor delineation: the weakest link in the search for accuracy in radiotherapy. J Med Phys Assoc Med Phys India 2008;33:136–40. https://doi.org/10.4103/0971-6203.44472.

[3] Das IJ, Compton JJ, Bajaj A, Johnstone PA. Intra- and inter-physician variability in target volume delineation in radiation therapy. J Radiat Res (Tokyo) 2021;62:1083–9. https://doi.org/10.1093/jrr/rrab080.

[4] Samarasinghe G, Jameson M, Vinod S, Field M, Dowling J, Sowmya A, et al. Deep learning for segmentation in radiation therapy planning: a review. J Med Imaging Radiat Oncol 2021;65:578–95. https://doi.org/10.1111/1754-9485.13286.

[5] 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.

[6] 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.

[7] Outeiral RR, 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.

[8] Ren J, Huynh B-N, Groendahl AR, Tomic O, Futsaether CM, Korreman SSPET. Normalizations to improve deep learning auto-segmentation of head and neck tumors in 3D PET/CT. In: Andrearczyk V, Oreiller V, Hatt M, Depeursinge A, editors. Head neck tumor segmentation outcome predict, vol. 13209. Cham: Springer International Publishing; 2022. p. 83–91. https://doi.org/10.1007/978-3-030-98253-9_7.

[9] Naser MA, Dijk LV van, He R, Wahid KA, Fuller CD. Tumor segmentation in patients with head and neck cancers using deep learning based-on multi-modality PET/CT images. 3D Head Neck Tumor Segmentation PETCT Chall., Springer; 2020, p. 85–98.

[10] Guo Z, Guo N, Gong K, Li Q, et al. Gross tumor volume segmentation for head and neck cancer radiotherapy using deep dense multi-modality network. Phys Med Biol 2019;64:205015.

[11] 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.

[12] Wong J, Fong A, McVicar N, Smith S, Giambattista J, Wells D, et al. Comparing deep learning-based auto-segmentation of organs at risk and clinical target volumes to expert inter-observer variability in radiotherapy planning. Radiother Oncol 2020;144:152–8.

[13] Rother C, Kolmogorov V, Blake A. “ GrabCut” interactive foreground extraction using iterated graph cuts. ACM Trans Graph TOG 2004;23:309–14.

[14] Castrejon L, Kundu K, Urtasun R, Fidler S. Annotating Object Instances with a Polygon-RNN. 2017 IEEE Conf. Comput. Vis. Pattern Recognit. CVPR, Honolulu, HI: IEEE; 2017, p. 4485–93. doi: 10.1109/CVPR.2017.477.

[15] Acuna D, Ling H, Kar A, Fidler S. Efficient Interactive Annotation of Segmentation Datasets with Polygon-RNN++. 2018 IEEECVF Conf. Comput. Vis. Pattern Recognit., Salt Lake City, UT: IEEE; 2018, p. 859–68. doi: 10.1109/CVPR.2018.00096.

[16] Boers T, Hu Y, Gibson E, Barratt D, Bonmati E, Krdzalic J, et al. Interactive 3D U-net for the segmentation of the pancreas in computed tomography scans. Phys Med Biol 2020;65:065002.

[17] Wang G, Li W, Zuluaga MA, Pratt R, Patel PA, Aertsen M, et al. Interactive medical image segmentation using deep learning with image-specific fine tuning. IEEE Trans Med Imaging 2018;37:1562–73.

[18] Smith AG, Petersen J, Terrones-Campos C, Berthelsen AK, Forbes NJ, Darkner S, et al. RootPainter3D: Interactive-machine-learning enables rapid and accurate contouring for radiotherapy. Med Phys 2022;49:461–73.

[19] Shahedi M, Halicek M, Dormer JD, Fei B. Incorporating minimal user input into deep learning based image segmentation. In: Landman BA, Išgum I, editors. Med. Imaging 2020 Image Process. Houston, United States: SPIE; 2020. p. 38. https://doi.org/10.1117/12.2549716.

[20] Klein S, Staring M, Murphy K, Viergever MA, Pluim JP. Elastix: a toolbox for intensity-based medical image registration. IEEE Trans Med Imaging 2009;29:196–205.

[21] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. Int. Conf. Med. Image Comput. Comput.-Assist. Interv., Springer; 2015, p. 234–41.

[22] Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-Net: learning dense volumetric segmentation from sparse annotation. Int. Conf. Med. Image Comput. Comput.-Assist. Interv., Springer; 2016, p. 424–32.

[23] Zhou Z, Rahman Siddiquee MM, Tajbakhsh N, Liang J. Unet++: A nested u-net architecture for medical image segmentation. Deep Learn. Med. Image Anal. Multimodal Learn. Clin. Decis. Support, Springer; 2018, p. 3–11.

[24] Zhou Z, Siddiquee MMR, Tajbakhsh N, Liang J. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. IEEE Trans Med Imaging 2019;39:1856–67.

[25] Li X, Sun X, Meng Y, Liang J, Wu F, Li J. Dice Loss for Data-imbalanced NLP Tasks 2020. doi: 10.48550/arXiv.1911.02855.

[26] Lin T-Y, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. Proc IEEE Int Conf Comput Vis 2017:2980–8.

[27] Yeung M, Sala E, Schönlieb C-B, Rundo L. Focus U-Net: A novel dual attention-gated CNN for polyp segmentation during colonoscopy. Comput Biol Med 2021;137:104815. https://doi.org/10.1016/j.compbiomed.2021.104815.

[28] 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.

[29] Wang G, Li W, Aertsen M, Deprest J, Ourselin S, Vercauteren T. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. Neurocomputing 2019;338:34–45. https://doi.org/10.1016/j.neucom.2019.01.103.