Physics and Imaging in Radiation Oncology 34 (2025) 100766

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

Enhancing patient-specific deep learning based segmentation for abdominal magnetic resonance imaging-guided radiation therapy: A framework conditioned on prior segmentation

Francesca De Benetti [a], Nikolaos Delopoulos [b], Claus Belka [b,c,d], Stefanie Corradini [b], Nassir Navab [a], Thomas Wendler [a,e,f], Shadi Albarqouni [g,h], Guillaume Landry [b], Christopher Kurz [b,*]

[a] Chair for Computer-Aided Medical Procedures and Augmented Reality, Technical University of Munich, Garching, 85748, Germany
[b] Department of Radiation Oncology, LMU University Hospital, LMU Munich, Munich, 81377, Germany
[c] German Cancer Consortium (DKTK), partner site Munich, a partnership between DKFZ and LMU University Hospital Munich, Munich, 81377, Germany
[d] Bavarian Cancer Research Center (BZKF), Munich, 81377, Germany
[e] Department of Diagnostic and Interventional Radiology and Neuroradiology, University Hospital Augsburg, Augsburg, 86156, Germany
[f] Institute of Digital Health, University Hospital Augsburg, Neusaess, 86356, Germany
[g] Clinic for Interventional and Diagnostic Radiology, University Hospital Bonn, Bonn, 53113, Germany
[h] Helmholtz AI, Helmholtz Munich, Oberschleißheim, 85764, Germany

ARTICLE INFO

Keywords:
MRI
Radiation therapy
MR-linac
Patient-specific segmentation

ABSTRACT

Background and purpose: Conventionally, the contours annotated during magnetic resonance-guided radiation therapy (MRgRT) planning are manually corrected during the RT fractions, which is a time-consuming task. Deep learning-based segmentation can be helpful, but the available patient-specific approaches require training at least one model per patient, which is computationally expensive. In this work, we introduced a novel framework that integrates fraction MR volumes and planning segmentation maps to generate robust fraction MR segmentations without the need for patient-specific retraining.

Materials and methods: The dataset included 69 patients (222 fraction MRs in total) treated with MRgRT for abdominal cancers with a 0.35 T MR-Linac, and annotations for eight clinically relevant abdominal structures (aorta, bowel, duodenum, left kidney, right kidney, liver, spinal canal and stomach). In the framework, we implemented two alternative models capable of generating patient-specific segmentations using the planning segmentation as prior information. The first one is a 3D UNet with dual-channel input (i.e. fraction MR and planning segmentation map) and the second one is a modified 3D UNet with double encoder for the same two inputs.

Results: On average, the two models with prior anatomical information outperformed the conventional population-based 3D UNet with an increase in Dice similarity coefficient > 4 %. In particular, the dual-channel input 3D UNet outperformed the one with double encoder, especially when the alignment between the two input channels is satisfactory.

Conclusion: The proposed workflow was able to generate accurate patient-specific segmentations while avoiding training one model per patient and allowing for a seamless integration into clinical practice.

1. Introduction

Linacs with an integrated magnetic resonance (MR) scanner (MR-linacs) have recently enabled MR-guided radiation therapy (MRgRT), a type of online adaptive RT, which has superior imaging quality compared to conventional cone-beam computed tomography-guided RT and allows real-time monitoring of the patient motion [1–3].

Conventionally, manual segmentations (“Segmentation-Planning”: S-P) based on the planning MR (“MR-Planning”: MR-P) are used, together with the deformably registered planning CT (“CT-Planning”: CT-P), to generate the RT treatment plan (RT-TP). The RT-TP is then updated before each RT session to account for changes in the anatomy

* Correspondence to: Marchioninistr. 15, 81377 Munich, Germany
E-mail address: christopher.kurz@med.uni-muenchen.de (C. Kurz).

https://doi.org/10.1016/j.phro.2025.100766
Received 30 August 2024; Received in revised form 2 April 2025; Accepted 5 April 2025
Available online 17 April 2025

and to reduce mistargeting of the RT. To do so, a fraction MR ("MR-Fraction": MR-Fx) is acquired and manually annotated (S-Fx). To reduce the workload of the manual annotation, MR-P is registered to MR-Fx, the same transformation is applied to S-P (S-P$^{\mathcal{R}}$) and used as guideline for S-Fx. Despite using S-P$^{\mathcal{R}}$, this is a time-consuming process, which can take up to ≈40% of the whole RT session [4].

The use of automatic segmentation algorithms based on deep learning (DL) would reduce the workload for the clinical team and the overall duration of the RT session [1]. However, most of these algorithms would produce a segmentation purely based on MR-Fx without taking into consideration the valuable information stored in S-P.

The problem of propagating the contours from the planning imaging to the fraction imaging or from previous fractions to following ones have been addressed in multiple ways.

A first set of works are based on the training of one model for each individual patient using only one volume and implemented different strategies to increase the generalizability of the model. Both Fransson et al. [5] and Li et al. [6] trained a 2D segmentation model per patient using only slices coming from the first fraction MR. In the former work, the segmentation of the later fractions was generated by ensembling the predictions of the last three best training checkpoints to reduce the misclassification errors [5], whereas in the latter the model was then finetuned with all the available fractions (i.e. from 1 to N-1) before segmenting fraction N [6]. Jeong et al. [7] trained a patient-specific (PS) segmentation model with one CT volume and tested on a second CT acquired at a later time point. As augmentation technique they generated 10 deformed CTs using Voxelmorph that were included in the training set.

Another set of work focused on finetuning population-based models to generate PS models. Chen et al. [8] first trained a population-based model with multi-patient first fraction MRs and then finetuned with the same strategy employed by Li et al. [6]. Similarly, Kawula et al. [9,10] used multi-patient planning MRs to train a population-based 3D UNet and then finetuned it using the planning MR of one specific patient, who was not included in the training set of the population-based model to generate a patient-specific UNet able to accurately segment the following fraction MRs of that specific patient. In both cases the finetuned PS models were shown to outperform the population-based model.

In general, the PS finetuning of baseline models, either population-based [8–10] or trained on a single volume [6], seems to be more robust than using models simply trained on one single volume without finetuning [5,7]. However, these works require the training of one model per patient, making them more difficult to use in clinical practice due to the larger computation resources and increased time required.

In this work, we present two DL-based approaches for models able to perform personalized segmentation taking advantage of prior anatomical information and without the need of having single PS models. Inspired by multi-modal image data fusion in segmentation tasks [11,12], our proposed approaches fuse the information of the fraction MR and the planning segmentation map to robustly predict the label map of the fraction MR.

2. Material and methods

We present a novel framework with two alternative models for PS multi-organ segmentation on MR-Fx: the first employs a nnUNet with dual-channel input (below referred to as 'nnUNet-DC'), and the second one employs a modified nnUNet with double encoder (nnUNet-DE). They take as input MR-Fx and S-P and extract relevant features from S-P to generate a more precise S-Fx. As baseline we used the default nnUNet set-up [13] (nnUNet).

All the approaches are in 3D and perform multi-organ segmentation of 8 abdominal regions of interest (ROIs) (see Section 2.1).

Table 1
Details of datasets: of the P patients included in the datasets (i.e. $P = 46$ in train/validation and $P = 23$ in test set), F_i of them fulfilled the inclusion requirements for MR-Fx i (i.e. the manual annotation of the 8 ROIs was available) and were included in the dataset (e.g. $F_1 = 44$ in the train/validation set). Therefore, the total number of volumes is given by the $\sum_{i=1}^{5} F_i$.

	Train/Validation	Test
Total patients	46	23
Fx 1	44	22
Fx 2	–	21
Fx 3	38	22
Fx 4	–	22
Fx 5	32	21
Total volumes	114	108

2.1. Dataset

The datasets consisted of 178 patients undergoing MRgRT for the treatment of various types of cancers in the abdomen at the LMU University Hospital (Munich, Germany) between January 2020 and November 2022. Informed written consent was obtained from all patients (LMU: ethics project number 20–291). A 0.35 T MR-Linac (MRIdian, ViewRay Inc, Cleveland, Ohio) was used to acquire MR volumes with a balanced steady-state free precession (bSSFP) sequence resulting in a T2*/T1 contrast. Each patient received between 1 and 20 RT fractions, and all the fractions were manually annotated by the clinical team during the RT session. The single-organ annotations were subsequently exported as point clouds and transformed into binary label maps using `plastimatch`. Finally, the single-organ annotations were merged into multi-organ volumetric label maps, in which each structure was assigned to an integer from 0 to N (where the background corresponds to label 0).

The training set was composed by MR-P, MR-Fx1, MR-Fx3 and MR-Fx5. Similarly, the test set included MR-P, MR-Fx1, MR-Fx2, MR-Fx3 MR-Fx4 and MR-Fx5. Patients with one or more missing MR-Fx among the required ones were included. Patients without MR-P (10 in total) or without MR-Fx (12 in total) were excluded. We considered as ROIs 8 structures that are clinically relevant in the treatment of various types of abdominal cancers and therefore whose annotations are abundant in the clinical data, namely: aorta, bowel, duodenum, kidney, right kidney, liver, spinal canal and stomach. We included in the dataset only MR-Fx in which the ground truth of all 8 ROIs were available. Patients with missing ground truth annotations in MR-P were excluded. The final training set consisted of 46 patients and 114 volumes, whereas in the test set there were 23 patients and 108 volumes (see Table 1).

The volumes in the training and test sets had variable voxel size (with the smallest being $1.49 \times 1.49 \times 3$ mm and the largest $1.63 \times 1.63 \times 3$ mm) and volume sizes (e.g $234 \times 234 \times 144, 266 \times 266 \times 144, 276 \times 276 \times 80, 360 \times 310 \times 144$ voxels). Further details about the data preprocessing can be found in Section 2.6.

2.2. Preliminary registration

The proposed methods make use of S-P as additional input. Similarly to the current clinical workflow, we ensured that S-P was registered to MR-Fx before feeding the inputs to the models. To do so, we first harmonized the spacing and the volume size, by resampling and zero-padding MR-P and S-P to match the spacing and the volume size of MR-Fx and S-Fx. After this, we rigidly registered MR-P to MR-Fx using `SimpleITK`, with the Normalized Cross Correlation as optimization metric. The transformation was then applied to S-P, thus generating S-P$^{\mathcal{R}}$, to bring it in the same coordinate space as MR-Fx.

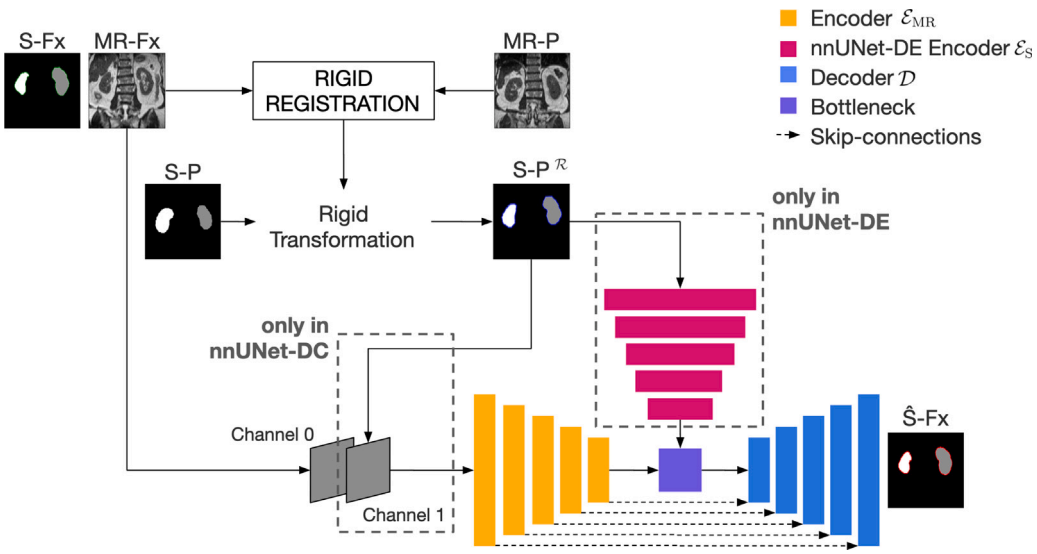


Fig. 1. Graphical representations of the architectures employed in this work. In the dashed boxes, the modifications to the baseline nnUNet architecture are specified: in the case of nnUNet-DC the input is dual-channel with MR-Fx and S-P$^{\mathcal{R}}$, whereas for nnUNet-DE, the first encoder $\mathcal{E}_{MR}$ processes MR-Fx and the second one $\mathcal{E}_S$ is used to process S-P$^{\mathcal{R}}$. For the sake of simplicity, only the kidneys are showed in the workflow.

2.3. nnUNet

For the baseline, we used the default 3D UNet [14] architecture[1] proposed by nnUNet [13], with an encoder ($\mathcal{E}$)-decoder ($\mathcal{D}$) structure and skip-connections. The input of this architecture is MR-Fx and the output is the segmentation of the 8 above-mentioned ROIs. The architecture has 6 stages with 32, 64, 128, 256, 320, 320 feature maps per stage, respectively. A $3 \times 3 \times 3$ kernel is used with stride of size [2, 2, 2] (with the exception of the first and last stages, where the strides are [1, 1, 1] and [1, 2, 2], respectively).

2.4. nnUNet with dual-channel input (nnUNet-DC)

The first of the proposed methods used the same architecture as nnUNet. The only difference is the input, which, in this case, is a dual-channel tensor with the first channel being MR-Fx and the second one S-P$^{\mathcal{R}}$.

2.5. nnUNet with double encoder (nnUNet-DE)

The backbone of the architecture of nnUNet-DE is nnUNet, therefore it was possible to use the same dynamically configurable architecture as for nnUNet and nnUNet-DC (see Section 2.3), which ensured the comparability of the three methods.

The architecture of nnUNet-DE is based on CloverNet [15], which consists in a 3D UNet, with encoder $\mathcal{E}_{MR}$, decoder $\mathcal{D}$ and an additional encoder $\mathcal{E}_S$, (see Fig. 1). The architecture of $\mathcal{E}_S$ and $\mathcal{E}_{MR}$ is exactly the same as $\mathcal{E}$ in nnUNet and nnUNet-DC. Similarly, the decoder $\mathcal{D}$ of the three architectures is the same. The input of $\mathcal{E}_{MR}$ is MR-Fx, hence the subscript "MR", and the input of $\mathcal{E}_S$ is S-P$^{\mathcal{R}}$, hence the subscript "S". The output is the multi-organ label map.

The latent space of $\mathcal{E}_S$ is concatenated to the latent space of $\mathcal{E}_{MR}$, fed to the bottleneck and afterwards to one common decoder $\mathcal{D}$. Due to the feature concatenation, the convolutional layer in the bottleneck of nnUNet-DE receives as input 640 feature maps (as opposed to the 320 in nnUNet) and returns 320 (same as in nnUNet). The skip connections, implemented via convolutions, are present only between $\mathcal{E}_{MR}$ and $\mathcal{D}$, as in the conventional UNet [13].

2.6. Training set-up

The three approaches were implemented in the widely adopted nnUNet framework [13], using the automatic nnUNet fingerprinting, network definition and training settings (see Table S1 in the Supplementary Material). The networks were trained with deep supervision for 1000 epochs, with an initial learning rate of 0.01, which decreased with the poly learning rate schedule [16], a combination of Dice Loss and Cross Entropy as loss function and the manual segmentation as ground truth, as proposed by the nnUNet framework. Similarly, the default nnUNet preprocessing (see Table S1) and data augmentation (including rotations, scaling, Gaussian noise, Gaussian blur, brightness, contrast, simulation of low resolution, gamma correction and mirroring) were applied.

Considering the 114 volumes in the train/validation set, they were split in 5 folds to perform cross-validation. To avoid data leakage between training and validation set, the default nnUNet random generation of the folds was not used. Instead, all the MR-Fx volumes belonging to one patient were manually assigned to one specific fold. As reported in Table 1, not all the patients in the train/validation set had all the fractions among MR-Fx1, MR-Fx3 and MR-Fx5, therefore the size of the validation set was variable across the folds ($\in$ [19, 25]).

At inference, one prediction per training fold was generated using a sliding window approach (with a window size equal to the size of the training patches), the final S-Fx was generated by averaging softmax probabilities generated by the 5 models (i.e. one per fold) and the default largest connected component postprocessing was applied.

2.7. Metrics

As evaluation metrics, we compared the predictions with the clinically used contours using the Dice similarity coefficient (DSC [%]), the 95th percentile Hausdorff distance (HD_{95} [mm]), and the normalized surface Dice (NSD [%], with thresholds set to 1 voxel), as implemented in the `monai` package.

The statistical significance of the difference between the approaches presented above per metric per organ was reported using the p-value[2] computed using the Wilcoxon signed-rank test, as implemented in the `scipy` package.

[1] The default version of 3D UNet in the nnUNet framework at the time of the development of this work was the one without the Residual Encoder preset.

[2] A p-value $\leq$ 0.05 indicates that the difference between the metrics reported by the two methods is statistically significant.

3. Results

The naive propagation of the contours of S-P to MR-Fx resulted in a median DSC over all the ROIs ($\overline{DSC}^{ROI}$) in the test dataset of 28% before the rigid registration (RR), whereas after the RR the $\overline{DSC}^{ROI}$ was 75%. The best alignment was found in terms of DSC in the kidneys and the liver ($\overline{DSC}$ > 83%), in terms of NSD and HD_{95} in aorta and spinal canal ($\overline{NSD}$ > 65%, $\overline{HD}_{95}$ = 5.1 mm). In terms of DSC, the worst was the duodenum ($\overline{DSC}$ = 56%) and, the bowel and stomach were the worst in terms of NSD and HD_{95} ($\overline{NSD}$ < 36%, $\overline{HD}_{95}$ > 12.5 mm). The quantitative analysis is reported in Table 2.

The nnUNet outperformed the RR with $\overline{DSC}^{ROI}$ = 81% and $\overline{NDS}^{ROI}$ = 70%, but achieved only $\overline{HD}_{95}^{ROI}$ = 7.8 mm.

Both nnUNet-DC and nnUNet-DE outperformed the nnUNet with an increase in $\overline{DSC}^{ROI}$ and $\overline{NSD}^{ROI}$ > 4%, and a decrease in $\overline{HD}_{95}^{ROI}$ > 2.5 mm. The difference between the performance of nnUNet and nnUNet-DC is significant ($p < 0.05$) for all the organs and all the metrics, excluding the DSC and HD_{95} of bowel. A similar behavior was observed for nnUNet-DE (see Table 2 for the details on statistical significance).

The largest improvement between nnUNet and nnUNet-DC was observed for aorta, bowel and spinal canal (≈9% in $\overline{DSC}$ and >9% in $\overline{NSD}$). In terms of $\overline{HD}_{95}$, the largest improvement was in the aorta (≈8 mm $\overline{HD}_{95}$).

On average, the performance of nnUNet-DE was worse than nnUNet-DC. Only in the case of the stomach, the $\overline{DSC}$ of nnUNet-DE was sightly better than the one of nnUNet-DC. On the other hand, nnUNet-DE performs better or comparable to nnUNet in all the ROIs.

The performance of nnUNet, nnUNet-DC and nnUNet-DE was similar for kidneys, liver and stomach with ≈1% $\overline{DSC}$ difference across the methods. A visual representation of the results is reported in Fig. 2.

The training was performed on a NVIDIA Quadro RTX 8000 GPU and took < 20 h for nnUNet and nnUNet-DC and ≈26 h for nnUNet-DE per fold. The difference in training time was expected given the different number of parameters of the architectures (≈30.8 million parameters in nnUNet and nnUNet-DC, and ≈56.2 million parameters in nnUNet-DE). It should be noted that the number of parameters is specific for the dataset presented above, as the three architectures considered in this work are dynamically built by the nnUNet framework depending on the characteristics of the dataset (e.g. spacing and image shape). At inference, the prediction of one volume took ≈30 s for nnUNet and nnUNet-DC and ≈47 s for nnUNet-DE.

Table 2
Evaluation with median and (interquartile range) of the presented approaches on the test set. The best metrics per organ are reported in bold. The statistical significance of nnUNet-DC and nnUNet-DE against nnUNet computed using the Wilcoxon signed-rank test (*: p-value ≤ 0.05, **: p-value ≤ 0.001) per organ is also reported.

	After RR	nnUNet	nnUNet-DC	nnUNet-DE
	DSC [%] NSD [%] HD_{95} [mm]	DSC [%] NSD [%] HD_{95} [mm]	DSC [%] NSD [%] HD_{95} [mm]	DSC [%] NSD [%] HD_{95} [mm]
Aorta	80 (10) 65 (24) 5.1 (2.7)	78 (11) 72 (12) 13.2 (15.5)	**87 (5)** ** **85 (11)** ** **4.8 (3.0)** *	85 (7) 82 (13) 5.0 (5.4) *
Bowel	69 (10) 35 (9) 15.0 (7.3)	72 (14) 44 (14) 19.9 (11.5)	**82 (9)** **53 (12)** ** **14.6 (7.7)**	80 (8) 52 (10) 16.1 (9.9)
Duodenum	56 (22) 41 (18) 9.5 (6.2)	66 (20) 56 (25) 16.0 (10.7)	**72 (16)** ** **62 (22)** ** **10.4 (12.1)** *	67 (20) 58 (25) 14.3 (12.5)
Kidney left	83 (12) 50 (30) 5.8 (2.9)	93 (5) 85 (17) 5.3 (4.7)	**94 (4)** ** **89 (17)** ** **3.7 (3.5)** **	93 (5) * 85 (17) * 5.1 (4.0)
Kidney right	83 (10) 51 (32) 5.4 (2.3)	92 (7) 82 (21) 7.1 (4.3)	**93 (7)** ** **86 (17)** ** **4.5 (4.0)** **	92 (7) 82 (21) 6.9 (3.8)
Liver	89 (5) 48 (24) 9.6 (6.1)	93 (3) 70 (13) 10.9 (10.9)	**95 (3)** ** **78 (15)** ** **9.7 (9.5)** **	94 (2) ** 74 (14) ** 10.2 (10.6) *
Spinal canal	75 (17) 67 (32) 5.1 (3.6)	79 (7) 77 (10) 10.7 (7.7)	**88 (6)** ** **90 (7)** ** **4.5 (3.0)** **	85 (5) * 86 (8) * 5.0 (3.5)
Stomach	67 (15) 36 (18) 12.5 (9.0)	88 (7) 77 (16) 10.1 (8.5)	89 (9) ** **80 (19)** ** **8.4 (7.9)** **	**89 (6)** ** 79 (15) ** 9.1 (7.1)
Average	75 (8) 50 (15) 5.3 (2.0)	81 (7) 70 (9) 7.8 (4.4)	**86 (5)** **77 (9)** **3.9 (2.0)**	85 (6) 74 (8) 5.1 (2.3)

4. Discussion

We implemented two different approaches which include the S-P in the learning process for the segmentation of MR-Fx and proved their efficacy on 8 clinically relevant ROIs in the abdominal region. In the first approach, we employed a nnUNet with dual-channel input (nnUNet-DC) and in the second one we used a modified nnUNet with an additional encoder (nnUNet-DE). We showed that, in general, the best way to introduce S-P is simply to add it as an additional channel to the encoder, after rigidly registering it to MR-Fx. In this way, S-P guides the segmentation of MR-Fx by focusing it on the correct areas.

Conventional population-based approaches for DL-based segmentation of different abdominal structures in multi-sequence MR datasets [17–21] reported results comparable to our nnUNet ones. Small differences in the metrics can be attributed to the different training set characteristics, such as number of samples and MR sequences, as well as the definition of the ROIs.

All the models presented in this work returned comparatively large interquartile ranges, which were on average smaller for nnUNet-DC and nnUNet-DE with respect to the baseline. Moreover, in few cases some segmentations resulted in DSC and NDS equal to zero (see Table 2). Both these aspects can be explained by the quality of the ground-truth segmentations in the dataset, which were acquired directly from the clinical workflow. Indeed, during RT, the annotation focuses on the relevant structures close to the tumor and is less accurate in the non-critical regions further from the tumor. An example can be seen in Fig. 3 (bottom).

In general, we observed that for ROIs in which the nnUNet was already >85% $\overline{DSC}$ (namely kidneys, liver and stomach), the contribution of S-P was limited. In all the other cases, the improvement when using nnUNet-DC was statistically significant in at least one of the metrics that we reported.

Most of the works on PS segmentation on the pelvic region [5,7,8] making the comparison with our proposed approach difficult. The proposed model of Li et al. [6] (namely, a single-patient model trained on MR-Fx1 and finetuned on MR-Fx2 to MR-Fx4 and tested on MR-Fx5) for the segmentation of various structures in abdominal, thoracic and pelvic regions returned better results than our proposed nnUNet-DC, both in terms of DSC and HD_{95}. However, they used T2-weighted MR and only evaluated their results in 6 patients. This could also explain their small standard deviations (≈1% in DSC and >4 mm in HD_{95}). Moreover, the label maps they used were generated by consensus of three senior radiation oncologists, with a final revision by physician director. It is therefore reasonable to assume that the quality of their annotation was higher than ours.

Kawula et al. initially applied their methodology to the pelvic region [9,10], but now are extending their analysis to the abdomen [22]. We extracted a subset of the test set presented above in order to have an additional test set that matched to the one used in their latest work, including only MR-P and MR-Fx5 of patients with all fractions from MR-Fx1 to MR-Fx5, for a total of 19 patients (two patients were excluded due to missing annotations). Their population-based model, trained on

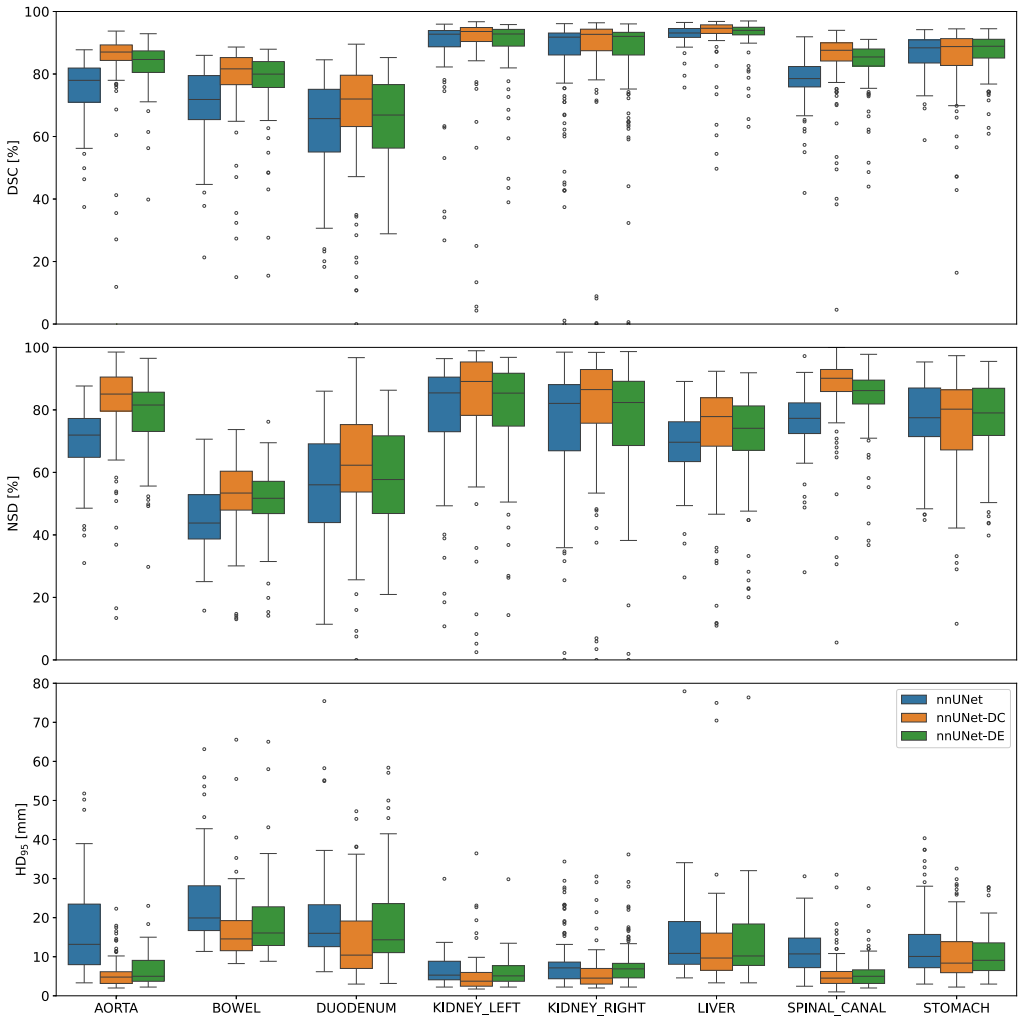


Fig. 2. Boxplot visualization of the metrics on the test set: DSC (top), NSD (mid) and HD_{95} (bottom). To improve the visualization of HD_{95}, some outliers ∈ [120, 140] mm were removed. The box extends from the first quartile (Q1 — 25th percentile) to the third quartile (Q3 — 75th percentile) of the data, with a line at the median. The whiskers extend from the box to Q3 — 1.5 IRQ and to Q1 + 1.5 IQR, where IQR = Q3 — Q1.

MR-P, was finetuned on MR-P and on MR-Fx1 to MR-Fx4 of one specific patient who was not included in the training set of the population-based model, and tested on MR-Fx5. The performance was similar, but slightly better than nnUNet-DC, with an absolute difference of $\overline{DSC}$ ≈2% in all the ROIs (excluding the duodenum, where the absolute difference was $\overline{DSC}$ ≈6%).

Similarly to the clinical workflow, both proposed methods include an initial rigid registration step. We analyzed its impact by generating an additional test set, with the same patients and fractions as the test set described in Table 1, but without applying the rigid registration. In general, Fig. 4 illustrates that the majority of organs with bad initial alignment (DSC < 60%) resulted in not satisfactory prediction (DSC < 60%) when nnUNet-DC was used.

Specifically, nnUNet-DC ($\overline{DSC}^{ROI}$ = 86%) was slightly superior to nnUNet-DE ($\overline{DSC}^{ROI}$ = 85%) when using the registration, and both nnUNet-DC and nnUNet-DE outperformed the nnUNet ($\overline{DSC}^{ROI}$ = 81%). However, when no registration was applied, the performance of nnUNet-DC dropped to $\overline{DSC}^{ROI}$ = 63%. Interestingly, for nnUNet-DE the drop was much smaller, being ≈5% in $\overline{DSC}^{ROI}$, with $\overline{DSC}^{ROI}$ = 80%. In this case, the performance of nnUNet-DE is slightly superior to nnUNet for aorta, bowel and stomach, whereas for the other organs both performed similarly. Similar behaviors were observed for HD_{95} and NSD (see Table S2).

We hypothesize that, when MR-Fx and S-P are misaligned, the additional encoder $\mathcal{E}_S$ of nnUNet-DE is able to extract more space invariant features compared to the encoder with dual-channel input of nnUNet-DC. This results in better performance for nnUNet-DE when the initial alignment is bad because the information of MR-Fx and S-P are combined in latent space, after the relevant features of each ones have been extracted separately. On the other hand, when the initial alignment is good, nnUNet-DC performs slightly better because it can match well the information stored in the two input channels by processing them together from the beginning.

Although nnUNet-DC and nnUNet-DE were slightly inferior to Li et al. [6] and Kawula et al. [9,10,22], they have higher clinical applicability because they require only one training phase, and the same model can be applied to multiple patients without the need of time- and memory-consuming PS finetuning.

To summarize, the proposed pipeline can be implemented with two alternative architectures and is a clinically integrated solution that incorporates PS data into radiation therapy using existing MRgRT workflows. It enhances personalized segmentation outcomes without disrupting clinical practice, allowing clinicians to speed up treatment planning and improve patient care without added complexity.

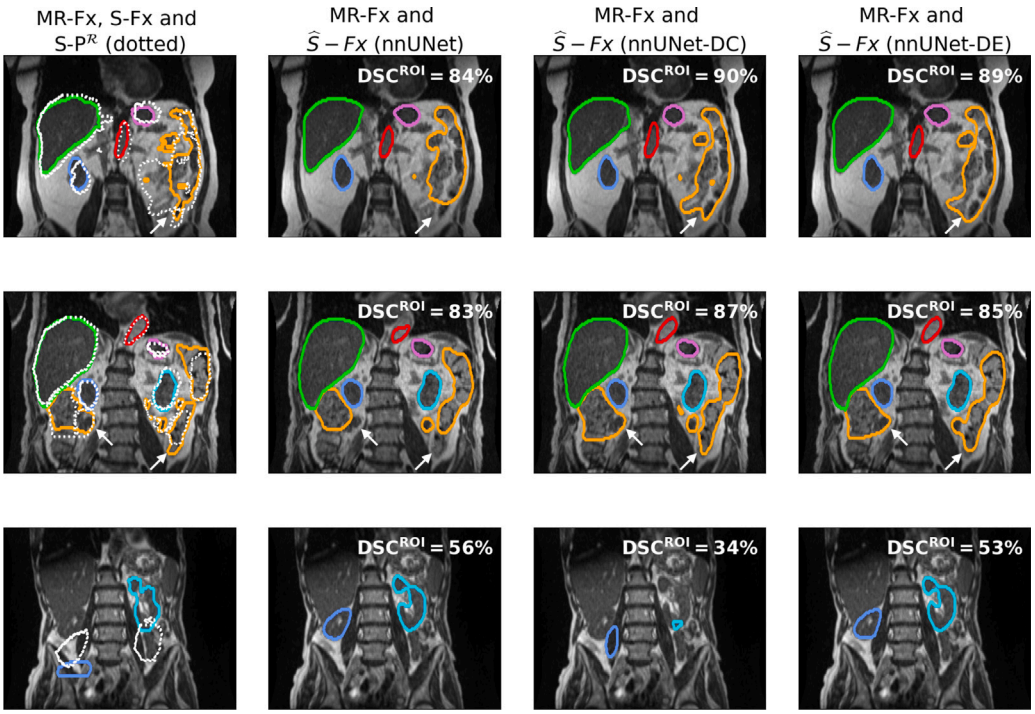
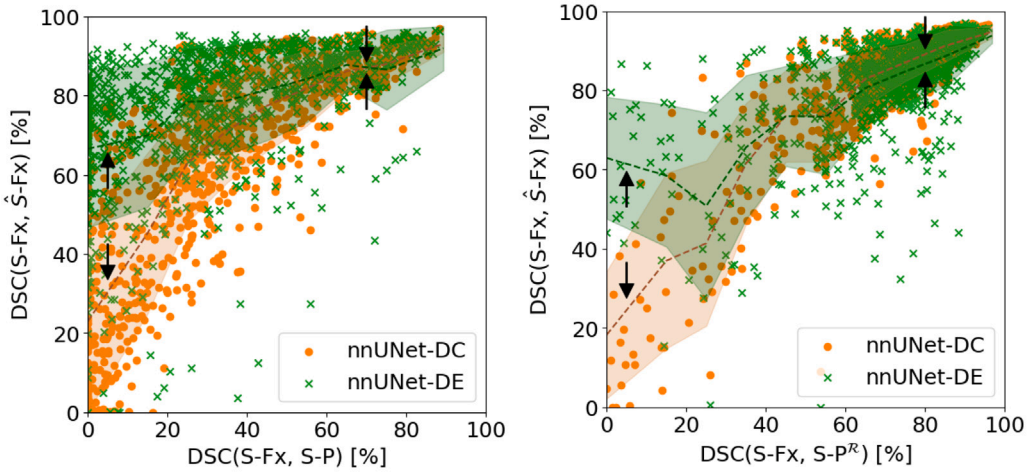


Fig. 3. Visual representation of the prediction of the three approaches. From left to right: MR-Fx with corresponding ground truth annotation (S-Fx) in color and planning annotation after RR (S-P$^{\mathcal{R}}$) in white (dotted); MR-Fx with corresponding predicted segmentation ($\hat{S}$-Fx) as generated by nnUNet; MR-Fx with corresponding $\hat{S}$-Fx as generated by nnUNet-DC; MR-Fx with corresponding $\hat{S}$-Fx as generated by nnUNet-DE. The ROIs are contoured in the following way: aorta — red; bowel — orange; right kidney — blue; left kidney — cyan; liver — green; stomach: purple. The average DSC of all the ROIs (DSC_{ROI}) is included.
Top and mid: coronal slices of two of the best performing samples.
Bottom: coronal slices of one sample with wrong S-Fx (see right kidney in blue) and failed RR. For the sake of clarity, only the kidneys are visualized. While the computation of the metrics is problematic because of the wrong S-Fx for all the approaches, the visual results are satisfactory for nnUNet and nnUNet-DE. For nnUNet-DC, the failed RR is the culprit behind the bad visual results. For a more detailed visualization of MR-Fx and S-Fx, please refer to Figure S1 in the Supplementary Material. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



(a) Without rigid registration (b) With rigid registration.

Fig. 4. Comparison of the initial DSC between the ground truth label map (S-Fx) and the planning label map (S-P), and the DSC between S-Fx and the predicted label map ($\hat{S}$-Fx). Each point/cross corresponds to one organ of one patient, the mean per bin (of size 10) is reported with the dashed line and the shadowed area represents ± one standard deviation from the mean. The black arrows point at regions in which the means of nnUNet-DC and nnUNet-DE are far apart (i.e. when the initial DSC is low) and when the means are close (i.e. when the initial DSC is high).

CRediT authorship contribution statement

Francesca De Benetti: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing – original draft, visualization. **Nikolaos Delopoulos:** Data curation, Writing – review & editing. **Claus Belka:** Data curation. **Stefanie Corradini:** Data curation. **Nassir Navab:** Conceptualization, Resources, Supervision, Funding acquisition. **Thomas Wendler:** Conceptualization, Supervision. **Shadi Albarqouni:** Conceptualization, Formal analysis, Writing – review & editing, Funding acquisition. **Guillaume Landry:** Conceptualization, Formal analysis, Resources, Data curation, Writing – review & editing, Funding acquisition. **Christopher**

Kurz: Conceptualization, Formal analysis, Resources, Data curation, Writing – review & editing, Funding acquisition.

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. The Department of Radiation Oncology of the LMU University Hospital has a research agreement with Elekta (Stockholm, Sweden) and Brainlab (Munich, Germany). Elekta did not fund this study, was not involved, and had no influence on the study design, the collection or analysis of data, or on the writing of the manuscript. Brainlab did not fund this study, was not involved, and had no influence on the study design, the collection or analysis of data, or on the writing of the manuscript.

Acknowledgment

This work was funded by the German Research Foundation (DFG, grant 469106425).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.phro.2025.100766.

References

[1] Guckenberger M, Andratschke N, Chung C, Fuller D, Tanadini-Lang S, Jaffray DA. The future of MR-guided radiation therapy. Semin Radiat Oncol 2024;34(1):135–44. http://dx.doi.org/10.1016/j.semradonc.2023.10.015.

[2] Ng J, Gregucci F, Pennell RT, Nagar H, Golden EB, Knisely JP, Sanfilippo NJ, Formenti SC. MRI-LINAC: A transformative technology in radiation oncology. Front Oncol 2023;13:1117874. http://dx.doi.org/10.3389/fonc.2023.1117874.

[3] Tchelebi LT, Zaorsky NG, Rosenberg J, Latifi K, Hoffe S. Integrating MR-guided radiation therapy into clinical practice: clinical advantages and practical limitations. Lung Cancer 2021;9:1289. http://dx.doi.org/10.37549/ARO1289.

[4] Landry G, Kurz C, Traverso A. The role of artificial intelligence in radiotherapy clinical practice. BJR| Open 2023;5(1):20230030. http://dx.doi.org/10.1259/bjro.20230030.

[5] Fransson S, Tilly D, Strand R. Patient specific deep learning based segmentation for magnetic resonance guided prostate radiotherapy. Phys Imaging Radiat Oncol 2022;23:38–42. http://dx.doi.org/10.1016/j.phro.2022.06.001.

[6] Li Z, Zhang W, Li B, Zhu J, Peng Y, Li C, Zhu J, Zhou Q, Yin Y. Patient-specific daily updated deep learning auto-segmentation for MRI-guided adaptive radiotherapy. Radiother Oncol 2022;177:222–30. http://dx.doi.org/10.1016/j.radonc.2022.11.004.

[7] Jeong S, Cheon W, Kim S, Park W, Han Y. Deep-learning-based segmentation using individual patient data on prostate cancer radiation therapy. PLoS One 2024;19(7):e0308181. http://dx.doi.org/10.1371/journal.pone.0308181.

[8] Chen X, Ma X, Yan X, Luo F, Yang S, Wang Z, Wu R, Wang J, Lu N, Bi N, et al. Personalized auto-segmentation for magnetic resonance imaging–guided adaptive radiotherapy of prostate cancer. Med Phys 2022;49(8):4971–9. http://dx.doi.org/10.1002/mp.15793.

[9] Kawula M, Hadi I, Nierer L, Vagni M, Cusumano D, Boldrini L, Placidi L, Corradini S, Belka C, Landry G, et al. Patient-specific transfer learning for auto-segmentation in adaptive 0.35 T MRgRT of prostate cancer: a bi-centric evaluation. Med Phys 2023;50(3):1573–85. http://dx.doi.org/10.1002/mp.16056.

[10] Kawula M, Vagni M, Cusumano D, Boldrini L, Placidi L, Corradini S, Belka C, Landry G, Kurz C. Prior knowledge based deep learning auto-segmentation in magnetic resonance imaging-guided radiotherapy of prostate cancer. Phys Imaging Radiat Oncol 2023;28:100498. http://dx.doi.org/10.1016/j.phro.2023.100498.

[11] Zhang Y, Yang J, Tian J, Shi Z, Zhong C, Zhang Y, He Z. Modality-aware mutual learning for multi-modal medical image segmentation. In: Medical image computing and computer assisted intervention–MICCAI 2021: 24th international conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I. Vol. 24, Springer; 2021, p. 589–99. http://dx.doi.org/10.1007/978-3-030-87193-2_56.

[12] Guo Z, Li X, Huang H, Guo N, Li Q. Medical image segmentation based on multi-modal convolutional neural network: Study on image fusion schemes. In: 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018). IEEE; 2018, p. 903–7. http://dx.doi.org/10.1109/ISBI.2018.8363717.

[13] Isensee F, Jaeger PF, Kohl SA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods 2021;18(2):203–11. http://dx.doi.org/10.1038/s41592-020-01008-z.

[14] Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: Medical image computing and computer-assisted intervention–MICCAI 2016: 19th international conference, Athens, Greece, 2016, Proceedings, Part II. Vol. 19, 2016, p. 424–32.

[15] De Benetti F, Yaganeh Y, Belka C, Corradini S, Navab N, Kurz C, Landry G, Albarqouni S, Wendler T. CloverNet–leveraging planning annotations for enhanced procedural MR segmentation: An application to adaptive radiation therapy. Work Clin Image- Based Proced 2024;1–10. http://dx.doi.org/10.1007/978-3-031-73083-2_1.

[16] Chen L-C, Papandreou G, Kokkinos I, Murphy K, Yuille AL. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE PAMI 2017;40(4):834–48. http://dx.doi.org/10.1109/TPAMI.2017.2699184.

[17] Akinci D'Antonoli T, Berger LK, Indrakanti AK, Vishwanathan N, Weiß J, Jung M, Berkarda Z, et al. TotalSegmentator MRI: Sequence-independent segmentation of 59 anatomical structures in MR images. 2024, ArXiv E-Prints.

[18] Häntze H, Xu L, Dorfner FJ, Donle L, Truhn D, Aerts H, Prokop M, van Ginneken B, Hering A, Adams LC, et al. MRSegmentator: Robust multi-modality segmentation of 40 classes in MRI and CT sequences. 2024, arXiv preprint arXiv:2405.06463.

[19] Zhou Y, Lalande A, Chevalier C, Baude J, Aubignac L, Boudet J, Bessieres I. Deep learning application for abdominal organs segmentation on 0.35 T MR-Linac images. Front Oncol 2024;13:1285924. http://dx.doi.org/10.3389/fonc.2023.1285924.

[20] Liang F, Qian P, Su K-H, Baydoun A, Leisser A, Van Hedent S, Kuo J-W, Zhao K, Parikh P, Lu Y, et al. Abdominal, multi-organ, auto-contouring method for online adaptive magnetic resonance guided radiotherapy: An intelligent, multi-level fusion approach. J Med Artif Intell 2018;90:34–41. http://dx.doi.org/10.1016/j.artmed.2018.07.001.

[21] Fu Y, Mazur TR, Wu X, Liu S, Chang X, Lu Y, Li HH, Kim H, Roach MC, Henke L, et al. A novel MRI segmentation method using CNN-based correction network for MRI-guided adaptive radiotherapy. Med Phys 2018;45(11):5129–37. http://dx.doi.org/10.1002/mp.13221.

[22] Kawula M, Marschner S, Wei C, Ribeiro MF, Corradini S, Belka C, Landry G, Kurz C. Personalized deep learning auto-segmentation models for adaptive fractionated magnetic resonance-guided radiation therapy of the abdomen. Med Phys 2024. http://dx.doi.org/10.1002/mp.17580.