Ayton, Norris
Bolus_CBCT
Fraction 5
13 APR 2017 7:54 AM
Male
Approve
Points
Beams (2D)
Goals
Overall Gamma
DVH
Images
Event Settings
Beams (2D)
Delivered
Difference
Zoom
Reset
Relative Dose
Difference
Profile Y
Profile X

Received: 5 October 2021 | Revised: 14 November 2021 | Accepted: 26 November 2021

DOI: 10.1002/mp.15421

RESEARCH ARTICLE

MEDICAL PHYSICS

Improving predictive CTV segmentation on CT and CBCT for cervical cancer by diffeomorphic registration of a prior

Chris Beekman | Suzanne van Beek | Jikke Stam | Jan-Jakob Sonke | Peter Remeijer

Department of Radiation Oncology, The Netherlands Cancer Institute, Amsterdam, The Netherlands

Correspondence
Peter Remeijer, Department of Radiation Oncology, The Netherlands Cancer Institute, Antoni van Leeuwenhoek Hospital, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands.
Email: p.remeijer@nki.nl

Abstract

Purpose: Automatic cervix-uterus segmentation of the clinical target volume (CTV) on CT and cone-beam CT (CBCT) scans is challenged by the limited visibility and the non-anatomical definition of certain border regions. We study the potential performance gain of convolutional neural networks by regulating the segmentation predictions as diffeomorphic deformations of a segmentation prior.

Materials and Methods: We introduce a 3D convolutional neural network that segments the target scan by joint voxel-wise classification and the registration of a given prior. We compare this network to two other 3D baseline models: One treating segmentation as a classification problem (segmentation-only), the other as a registration problem (deformation-only). For reference and to highlight the benefits of a 3D model, these models are also benchmarked against a 2D segmentation model. Network performances are reported for CT and CBCT segmentation of the cervix-uterus CTV. We train the networks on the data of 84 patients. The prior is provided by the CTV segmentation of a planning CT. Repeat CT or CBCT scans constitute the target scans to be segmented.

Results: All 3D models outperformed the 2D segmentation model. For CT segmentation, combining classification and registration in the proposed joint model proved beneficial, achieving a Dice score of 0.87 and a mean squared error (MSE) of the surface distance below 1.7 mm. No such synergy was observed for CBCT segmentation, for which the joint and the deformation-only model performed similarly, achieving a Dice score of about 0.80 and an MSE surface distance of 2.5 mm. However, the segmentation-only model performed notably worse in this low contrast regime. Visual inspection revealed that this performance drop translated into geometric inconsistencies between the prior and target segmentation. Such inconsistencies were not observed for the deformation-based models.

Conclusion: Constraining the solution space of admissible segmentation predictions to those reachable by a diffeomorphic deformation of the prior proved beneficial as it improved geometric consistency. Especially for CBCT, with its poor soft-tissue contrast, this type of regularization becomes important as shown by quantitative and qualitative evaluation.

1 | INTRODUCTION

Accurate delivery of external beam radiotherapy of cervical cancer is challenged by a large day-to-day motion of the clinical target volume (CTV) of the cervix-uterus, mainly due to varying bladder and rectal filling.[1] While the library of plans methods can be used to mitigate the effect of the large anatomical deformations,[2,3] image-guided adaptive radiotherapy is increasingly moving toward a fully online scenario.[4] For treatment plan adaptation on the fly, fast re-contouring is a necessary hurdle to overcome. Deep neural networks provide state of art in fast auto-segmentation of medical images. However, the performance of computer vision tasks using neural networks is dependent on the input image quality.[5] The large interobserver delineation uncertainty of the cervix-uterus CTV and the need for clear delineation guidelines[6] indicate that its anatomical borders are not very distinct on CT. On CBCT, the borders are even less pronounced due to inferior image quality. Such ambiguity in the input data may not only hamper consistent manual segmentations, but also pose difficulties for automatic segmentation, making the cervix-uterus CTV a challenging target.

Furthermore, medical segmentations in radiotherapy may not only be a function of the input image, but may also depend on certain patient-specific clinical factors. For instance, additional information that can impact the delineation process may be obtained from gynecologic examination and other imaging modalities such as MRI or PET. Moreover, the definition of the cervix-uterus CTV is partially motivated by the clinical experience that has led to a varying extent of inclusion of the parametria and vagina.[7] Consequently, its borders do not strictly correspond to visible borders on medical images. While such delineation decisions should propagate throughout treatment, it is unlikely that a machine learning algorithm will reproduce them from the input image alone. This indicates the need for patient-specific prior information for accurate auto-segmentation of the CTV.

Segmentation problems are often tackled by training 2D or 3D U-Nets that perform a classification of the input scan. While 2D models are more computational and memory efficient, they lack contextual information. The segmentation network by Nikolov et al.[8] incorporates contextual information from the third direction in their architecture, yet predictions are still made slice by slice. Such intrinsic 2D models know very little about the underlying anatomy to be segmented, challenging consistent segmentation over the slices. On the other hand, 3D models have the potential of learning the concept of a 3D shape, along with desired properties such as smoothness and connectedness. Such behavior can be further encouraged by providing shape priors to the model, for example by conditional random fields,[9,10] by learning probabilistic anatomical priors,[11] or by explicitly providing a template segmentation as additional input.[12] A restrictive prior is given in the context of deformable templates, as in atlas-based methods. Here, target and atlas scans are registered in order to propagate the atlas segmentation to the target scan.[13] If the deformation field is diffeomorphic, that is, a bijective, smooth mapping with a smooth inverse, topological consistency with the template shape is guaranteed, in the sense that the template shape is smoothly deformed without merging or tearing.

This topological consistency can be beneficial since it is a minimal anatomical requirement, and effectively narrows down the solution space of possible model predictions. Models that treat segmentation as a classification problem lack this guarantee. On the other hand, a deformable template model comes with a number of drawbacks too. First, as deformed instances of the template, the predictions may be quite biased towards the provided template shape. It is therefore important that the template is a good representation of the possible target shapes. Defining a good template is a research question in its own right.[13] Second, registration has proven to be a harder problem than classification. While neural networks are considered state-of-the-art for classification,[14] their performance on registration problems is trailing, especially for large deformations. There may be various reasons for this. First, the number of degrees of freedom is much larger in registration than in classification. Furthermore, unsupervised registration is somewhat ill-posed in that there are many acceptable possible registrations that are similarly scored by the loss function, which makes gradient descent challenging.[15]

Joint learning of both segmentation and registration potentially leverages the strengths of both tasks. In particular, the classification part may guide the registration, while the registration ensures topologically valid anatomies. Furthermore, the advantage of joint learning could be more intrinsic: By learning a shared representation for multiple related tasks, the generalization properties of the model are potentially improved.[16] Joint registration and segmentation have been proposed in a more classical setting, as well as in deep learning approaches. Some classical methods realize image segmentation as the registration of a prior contour using variations of active contour methods.[17,18] Chen et al.[19] proposed a probabilistic process in which an initial segmentation estimate is continuously updated. Essentially, it finds the maximum a posteriori (MAP) estimators of the deformation field parameters given the current segmentation estimate followed by MAP estimation of the segmentation given the update transformation. This process is iterated until convergence is reached. Methods to simultaneously update the transformation and segmentation have also been proposed.[20] In this study, we draw inspiration from this idea to iteratively update the deformation and classification predictions together, but cast it into a neural network. This is in contrast to other deep learning approaches for joint registration and segmentation, which typically predict both the segmentation and deformation field in one go.[12,21–23]

Classical schemes typically re-optimize every image pair separately. Consequently, common relationships across the data are not discovered, and therefore not extrapolated to individual instances. Moreover, this individual iterative re-optimization renders such schemes computationally expensive. Deep

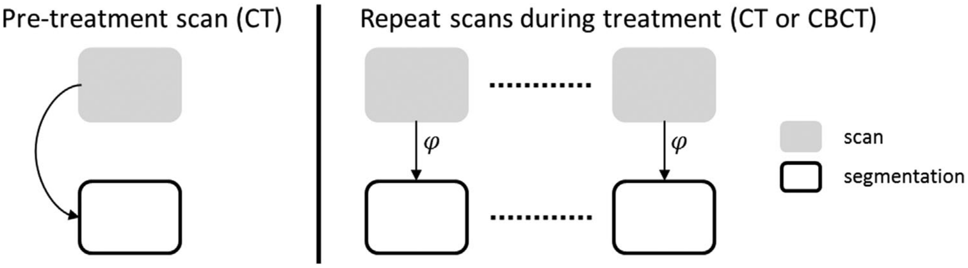

FIGURE 1 Schematic overview of a common radiotherapy scenario. The goal is to find a function φ that maps the target scan, given the prior information of the pre-treatment phase, to a segmentation

learning-based approaches try to overcome both issues by optimizing some appropriate loss over the entire training data set. Once trained, the inference is performed in a single forward pass, avoiding the costly re-optimization. The downside of this is that neural networks cannot recover from prediction errors as no feedback is provided during inference. For predicting large deformations as in our case, it has been shown that iteratively correcting the predicted deformation field enhances performance at the cost of longer inference time.[24–26] The iterative network we propose here naturally allows for this.

In this study, we trained three 3D CNNs for resegmentation on CT and CBCT, each using a segmentation prior. In particular, we proposed a joint registration and segmentation model and compared it to its separate components: (1) A segmentation-only model, treating the task as a classification problem; (2) A deformation-only model, predicting a diffeomorphic deformation of the given prior. Furthermore, we compared the 3D CNNs with a previously published 2D segmentation CNN for performance reference, and in order to highlight the advantages of 3D over 2D.

Our research objective was twofold: First, to investigate whether joint learning outperforms separate learning of classification and registration as a segmentation strategy; and second, to determine the effect of image quality on these different approaches. We carried out this study in the context of re-segmentation of the cervix-uterus CTV, where the segmentation on the planning CT is taken as the prior from which shape characteristics should be inferred. Both CT and CBCT segmentations were considered to study a real-world image quality deterioration scenario.

2 | MATERIALS AND METHODS

The real-world scenario we consider is schematically depicted in Figure 1. We look for a function φ that maps a repeat scan x (either CT or CBCT) to its CTV segmentation y given the prior information of the pre-treatment planning stage; the planning CT and its CTV segmentation $\{x^P, y^P\}$. The function φ will be parametrized by a CNN, as will be detailed below.

2.1 | Data and preprocessing

A group of 84 cervical cancer patients (IRB approved), treated at The Netherland Cancer Institute in recent years were randomly split into a training ($n = 64$), validation ($n = 10$), and test ($n = 10$) set. Two CT scans were available for each patient. One scan was made in full bladder state, for which the patient had received drinking instructions. The other scan was made directly after voiding of the bladder, hence in an empty bladder state. On each of these scans, the CTV of the cervix-uterus was manually delineated. Additionally, four daily CBCTs were delineated for a subgroup of 20 patients from the training set, and two for each patient in the validation and test sets. The CBCTs are routinely acquired as part of the clinical workflow, but normally not delineated. Hence, the 120 additional CBCT delineations were considered to be an acceptable tradeoff between workload and data quantity. For each patient, CT and CBCT scans were registered on bony anatomy. All scans were normalized by linearly rescaling between 0 and 1 after clipping between −300 and 500 Hounsfield units.

To ensure consistent positioning across patients, the scans and segmentations were shifted for each patient so that the center of mass of the CTV on full bladder CT coincided with the origin. Subsequently, the volumes were cropped by the bounding box of the union of all CTVs, expanded by a 5 cm margin. Finally, all volumes were resampled to a 128 × 80 × 128 dimensional grid in order for the model to fit into GPU memory, resulting in a voxel size of 0.185 × 0.3 × 0.186 cm^3. To obtain segmentations in the same grid as the original target scan, a linear resampling can be performed on the final segmentation probability map and deformation.

Paired data augmentation for the prior and target data was performed. Data augmentation included random shifts, rotations, scaling to improve robustness against small positional and scale changes; and flipping about the sagittal plane, assuming approximate left-right

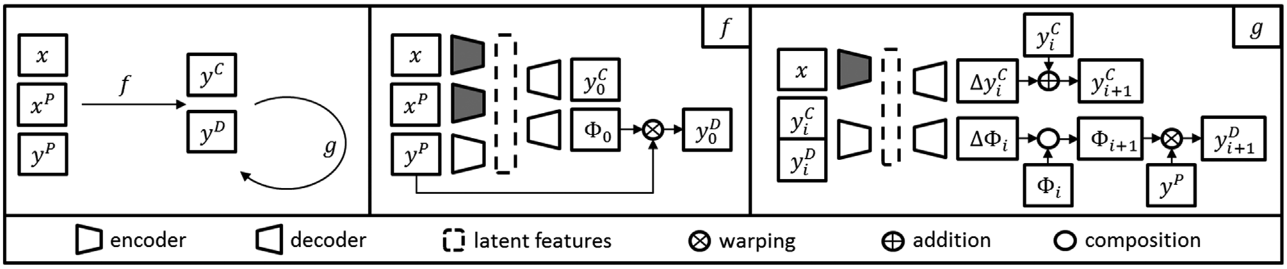

FIGURE 2 Overview of the proposed network. On the left: The initialization U-Net f uses a prior scan-segmentation pair $(x^P,\ y^P)$ to predict two initial segmentations for the target scan x: a voxel-wise classification map, y^C, and a deformation of the segmentation prior, y^D. The refinement U-Net g takes these predictions as input and updates them simultaneously. In the middle and right box, network layouts of f and g are depicted in more detail. Shaded encoders share parameters. Network g can be cascaded for iterative refinements

symmetry. Furthermore, the scan data were augmented by slight brightness and contrast changes. These augmentations were performed on the fly during training.

2.2 | Joint registration and segmentation model

In line with previous works, segmentations, and registrations were predicted from an input pair using a 3D CNN.[27,28] Moreover, following Zhu et al.,[21] we adopted the multi-task learning paradigm and learned the voxel-wise classification map and deformation from the same latent feature representation of the input. This shared latent representation was learned by training a single encoder for both tasks. From the latent encoding, two separate decoders map the latent representation to a segmentation prediction and a deformation field prediction that warps the prior segmentation. Hence, two predictions were made for the target scan; we let y^C denote the classification probability map and y^D the warped prior segmentation. Both y^C and y^D take values in the range $[0,\ 1]$.

While this network could be trained in a joint fashion using appropriate loss terms, the coupling between the segmentation and the deformation then only resides in the shared representation. To make this coupling more explicit, a refinement U-Net g is appended to an initialization U-Net f. The refinement network g takes the output of f as input along with the target scan and refines the initial prediction for the segmentation and deformation (see Figure 2). This serves two purposes:

1. Since the predictions made by the initial segmentation and deformation decoder are effectively fed back into the network, the exchange of information between the two tasks can occur during inference. As a consequence, the network can refine the segmentation prediction based on the deformation prediction and vice versa.
2. Observed deformations of the cervix-uterus CTV can be notably large, challenging accurate registrations. Cascaded architectures, in which deformation fields are iteratively refined, can recover from residual misalignments and thus improve registration.

2.3 | Baseline models

To examine the efficacy of the joint iterative learning approach, we compared it to a prior-aware 3D segmentation-only model, and a deformation-only model that deforms the given prior. In this segmentation-only model, only the segmentation map y^C is predicted by f. We removed the refinement network since the classification predictions are now solely based on the inputs, which are already present in f. On the other hand, for the deformation-only model, we kept the refinement network to iteratively refine the initial prediction of y^D.

Furthermore, the segmentation networks were benchmarked against a baseline provided by the established network architecture proposed by Nikolov et al.,[8] which was specifically fine-tuned for optimal performance on the considered task. In particular, class imbalance was a major problem in training this network due to the vast amount of empty slices. This is a direct consequence of the intrinsic 2D nature of the network. We used the class-balanced loss as proposed by Cui et al.[29] to resolve this issue.

2.4 | Diffeomorphic deformations

Diffeomorphic deformations were ensured by adopting the "stationary velocity field" (SVF) framework.[30] In this framework, deformations are obtained by flowing along a time-independent velocity field v for some time t. This flow is described by the following ordinary differential equation:

$$\frac{d\Phi(t,\ x)}{dt} = v(\Phi(t,\ x)) \tag{1}$$

A diffeomorphic deformation $\Phi^\tau(x) := \Phi(\tau, x)$ is obtained by time integration of Equation (1) between $t = 0$ and $t = \tau$ for $\tau \in \mathbb{R}$. In particular, Φ^0 is the identity transform and Φ^1 denotes the final deformation. Instead of predicting the deformation directly, an SVF v is predicted by the network, which is then integrated over time to yield a diffeomorphic deformation. We efficiently computed this integration using "squaring and scaling,"[31] as was previously done in a deep learning context by Dalca et al.[32] The smoothness of the deformations Φ^t is inherited from the generating velocity fields v. To ensure smoothness of the SVF, the network output was smoothed by an isotropic Gaussian kernel of size σ. This operation was performed in Fourier space. Henceforth, the superscript on the deformations will be dropped. We write Φ to mean Φ^1 since we are only interested in the final deformation.

2.5 | Network architecture

Although the convolutional sub-networks f and g operate on different inputs, their general architecture is the same. We separately encoded each of the inputs to map them to the latent space. From this latent space, two decoder arms originate: one to map the latent representation to a deformation used to warp the prior segmentation, the other to predict a segmentation directly.

2.5.1 | The encoder architecture

The role of the encoder is to extract relevant features from the input data. We used separate encoder networks for scans and segmentations. In particular, we used an encoder consisting of four residual blocks with max-pooling layers in between to extract relevant features from input scans. As segmentations contain fewer features than medical scans, we assumed that fewer filters were necessary for the segmentation encoder. Hence, for segmentation input, four convolutional blocks were used, containing fewer parameters. The assumption here is that relevant features are more easily extracted from segmentations than from scans. To further limit the number of model parameters, the same convolutional kernels were used in encoding the prior and target scan. However, the internal layer normalizations were modality-specific as proposed in Dou et al.[33] since the target scan might be either a CT scan or a CBCT scan. In this way, relevant features which generalize to CBCT scans can be learned by leveraging the superior image quality on CT. To this end, both CT and CBCT data were presented to the scan encoder during training of the 3D networks. The latent representation was then obtained by concatenating the extracted features from the scan and segmentation inputs at every level. The encoder architecture is depicted in Figure 3, where the number of filters used in each convolutional layer is indicated. For every convolutional layer, an isotropic kernel size of 3 was used.

2.5.2 | The decoder architecture

Two separate decoders map the latent representation to the network prediction. We will refer to these as the "deformation arm" and the "segmentation arm" of the network. In the deformation arm, the latent features were mapped to a dense vector field of the same dimension as the original inputs. This vector field is convolved with a Gaussian kernel to obtain a smooth velocity field. A kernel size of $\sigma = 5$ was empirically found to yield satisfactory results for our purpose. This velocity field is integrated into five scaling and squaring steps to yield the deformation used to warp the prior segmentation. In f, the initial velocity field v_0 is predicted and integrated to Φ_0, the initial deformation field. Based on the output of f, a refinement velocity field Δv_0 is predicted in g, and integrated to $\Delta\Phi_0$. The updated deformation field is then obtained as the composition $\Delta\Phi_0 \circ \Phi_0$. This process can be iterated for multiple refinement steps by cascading the refinement module (Figure 2), where the final registration field is continuously updated according to $\Phi_{i+1} = \Delta\Phi_i \circ \Phi_i$.

Simultaneously, a classification probability map is predicted in the segmentation arm. While in the initialization module a sigmoid activation function is used to squeeze the network output in the range $(0, 1)$, in the refinement model, we used a hyperbolic tangent function. This maps the network output ρ to the range $(-1, 1)$. The segmentation correction map at iteration i is then obtained as follows:

$$\Delta\, y_i^C = \begin{cases} \tanh(\rho) \cdot y_i^C \;\; where \;\; \tanh(\rho) \leq 0 \\ \tanh(\rho) \cdot \left(1 - y_i^C\right) \;\; where \;\; \tanh(\rho) > 0 \end{cases} \tag{2}$$

With this definition, the corrected segmentation map $y_{i+1}^C = y_i^C + \Delta y_i^C$ is ensured to remain in the range $(0, 1)$. Intuitively, we cannot "subtract" more probability than was originally there, and cannot "add" more probability than its complement admits. After the final iteration is performed, the conversion to a binary segmentation is performed in a post-processing step, by thresholding at 0.5.

Because of the iterative nature of the network, predictions made in the respective arms of the decoder can exchange information. This mutual guidance potentially allows the network to leverage the strengths of both tasks. The decoder architecture is shown in Figure 4. As before, the number of filters in each layer is indicated and an isotropic kernel size of 3 is used.

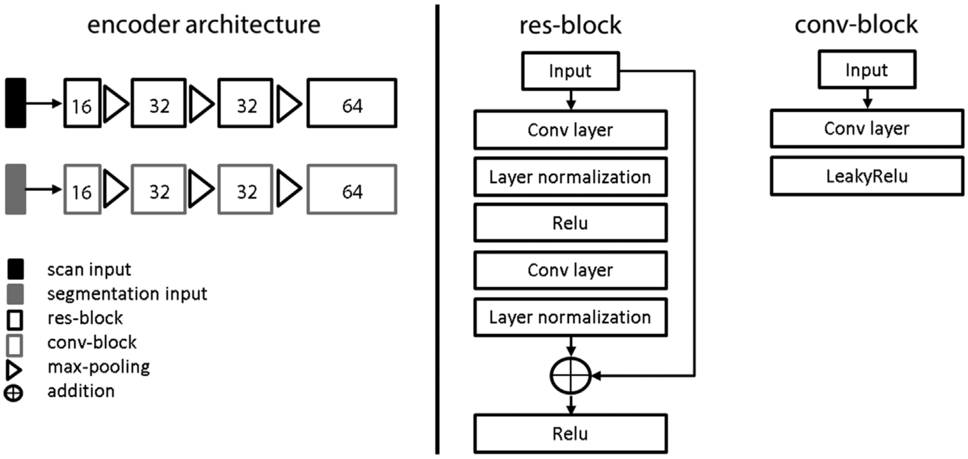

FIGURE 3 The encoder architecture is similar for scan and segmentation inputs (left), except for the feature extraction blocks (right). All operations were performed in 3D

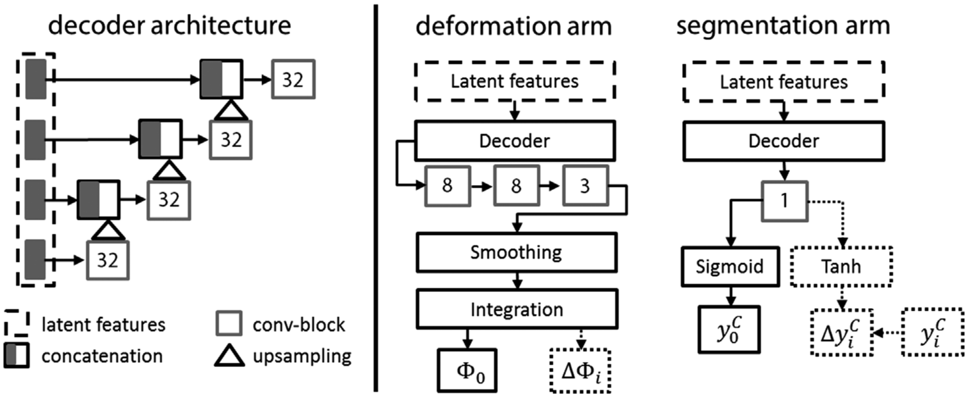

FIGURE 4 The decoder architecture (left). Separate decoders are trained for the deformation arms and the segmentation arms (right). Dotted paths are used in the refinement network *g*

2.6 | Loss functions

The loss was defined as $\mathcal{L} = \mathcal{L}_{\text{seg}} + \mathcal{L}_{\text{def}} + \mathcal{L}_{\text{norm}}$. The first term encouraged accurate prediction of the segmentation; the second term optimized deformation predictions. The last term ensured the prediction updates by the refinement network *g* are small, allowing for the iterative usage of it.

$\mathcal{L}_{seg}$ was defined as the binary cross-entropy between the true segmentation and the final predicted segmentation given by

$$\mathcal{L}_{\text{seg}} = -\left(ylog\left(y^C\right) + (1-y)\log\left(1-y^C\right)\right) \tag{3}$$

$\mathcal{L}_{def}$ was defined as a weighted sum of the Dice loss function and a regularization term:

$$\mathcal{L}_{def} = \mathcal{L}_{Dice} + \alpha\|\nabla\cdot v\|^2 \tag{4}$$

By minimizing the squared *l*2-norm of the divergence of the final velocity field *v*, volume-preserving deformation fields are favored. This reflects the clinical observation that the volume of the cervix-uterus CTV remains roughly equal. By adding this term to the loss function, the solution space of possible velocity fields is restricted to those that are roughly divergence-free, making the network potentially more robust against anatomically implausible predictions.

Finally, the $\mathcal{L}_{\text{norm}}$ term encourages the initial predictions to be globally accurate, such that refinements are small and localized. It does so by minimizing the squared *l*2-norms of the segmentation correction map Δy_0^C, and the refinement velocity field Δv_0. That is,

$$\mathcal{L}_{norm} = \beta\left\|\Delta y_0^C\right\|^2 + \gamma\|\Delta v_0\|^2 \tag{5}$$

Note that we train with only one refinement.

2.7 | Training details

The networks were implemented in Keras with Tensorflow backend, and trained on a Tesla K80 graphics card. We used the Adam-optimizer with a learning rate of 10^{-4}. Due to memory limitations, the batch size was restricted to 1; which is also the reason why we use layer normalization instead of the more common batch normalization. We trained the network for 350 epochs, with cyclic stochastic weight averaging used in the last 50 epochs in order to obtain a more robust model.[34] Good model performance was achieved with the weighting parameters set to $\alpha = 5, \beta = 1, \gamma = 0.5$. While the network was trained using one refinement pass to determine suitable network parameters for f and g, during inference we iterated the refinements.

2.8 | Network performance

The 3D networks were trained on CT and CBCT data jointly. The optimal number of refinement corrections during inference was determined by scoring the Dice score and the number of connected components against the number of refinements used. We report performances on CT and CBCT test data. The performance is evaluated in terms of four metrics: the Dice score; the mean squared error (MSE) surface distance; the 90 percentile surface distance; and the percentage of predictions in which the segmentation consisted of multiple connected components. We benchmarked the performance of the proposed joint model against the 2D segmentation model, the 3D segmentation-only model, and the 3D deformation-only model. For statistical comparison, the 3D segmentation-only and deformation-only models were taken as reference. *P*-values were obtained for the relative performance of the other models, using a two-sided Wilcoxon signed-rank test. We obtained multiple data points per patient: For CT, predictions were made for the full or empty bladder scan with the empty or full bladder anatomy used as prior, respectively. For CBCT, either CT scan was taken as prior and segmentation predictions were made for each of the available CBCTs. Since these data points are not independent, we averaged the scores for each patient and used these average scores as independent data points for the statistical tests. Finally, qualitative differences between predicted segmentations were highlighted based on visual observation of the worst, median, and best scoring segmentation instances of the 2D model.

3 | RESULTS

As previously mentioned, we trained the 3D networks using both CT and CBCT data, resulting in a single model used for both modalities. For completeness, we compared this approach to training on CT and CBCT data separately. While for CT segmentation, adding CBCT data to the training set did not noticeably change performance, the concurrently trained models performed much better on CBCT than the models trained on CBCT data only. This indicates that the quality of CBCT data alone was insufficient for the network to extract features that generalize well to unseen data. Results are therefore reported for the models trained concurrently on both modalities.

TABLE 1 Inference timings in seconds of the different 3D networks

Corrections	Base	1	3	5
Model				
2D	0.99	×	×	×
3D seg only	4.4	×	×	×
3D def only	7.9	12.2	18.0	24.1
3D joint	9.0	14.5	22.3	30.3

3.1 | Optimal number of refinement corrections

The left-hand side of Figure 5 shows the Dice scores of the predicted segmentation for CT and CBCT test data as the number of refinement corrections increases. Iterating the refinement module proved stable though showed little efficacy. Some improvement can however be achieved for individual cases. While the deformation arm of the network predicts segmentations as a single connected component by construction, the segmentation arm has no such guarantee. In fact, as can be observed from the RHS of Figure 5, initial predictions by the segmentation arm often result in multiple connected components, especially for CBCT. During the iterative refinements, however, the deformation arm provides guidance, encouraging the segmentation arm to predict a single connected component as well while maintaining accuracy. Moreover, while the base predictions from the segmentation and deformation arm typically differ, they quickly converge to agreement upon iterating the refinements yielding residual differences negligible. It should be noted that adding refinement corrections increases the inference time (Table 1). We used three refinement corrections for the remainder of this study, based on these observations.

3.2 | Network performance

Network performances are shown in Table 2. For the 2D model, obtaining consistent segmentation results on CT data proved challenging. Therefore, no attempt was made to conquer the even more challenging case of

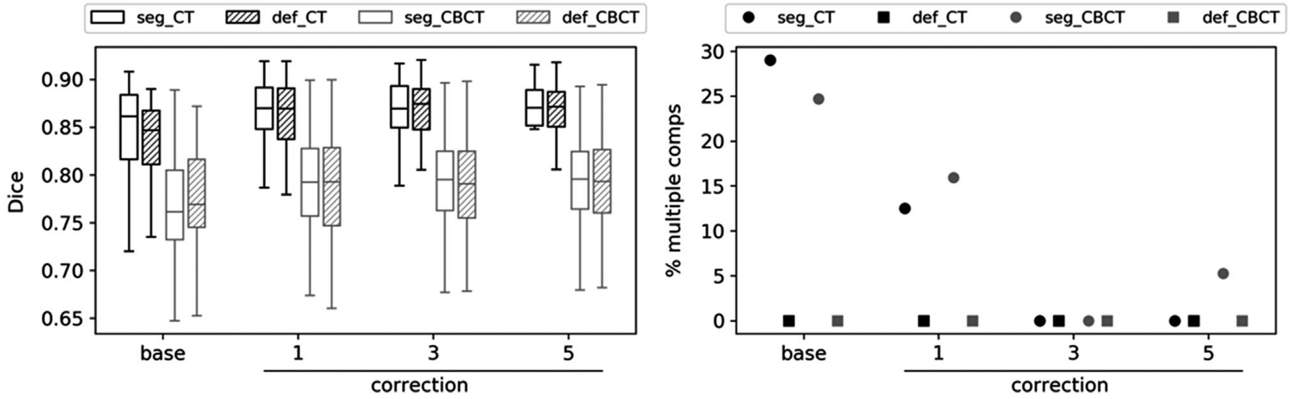

FIGURE 5 Trends in the CT and CBCT segmentations are predicted by the segmentation and the deformation arm of the network as the number of correction increases. **LHS**: Boxplot showing the Dice score of the predicted segmentations. **RHS**: Percentage of segmentation predictions with multiple connected components

TABLE 2 Network performances on the test set. Median values reported. Statistical tests were performed with respect to the 3D segmentation only model (superscripts) and the 3D deformation only model (subscripts). *P*-values below the 0.05-level are indicated: **p*≤0.05; ***p*≤0.01. The last column was not tested on significance as the percentage of predictions resulting in multiple components is a number, rather than a distribution

	Dice		MSE surf dist (cm)		90th surf dist (cm)		Multiple comps (%)	
Metric Model	CT	CBCT	CT	CBCT	CT	CBCT	CT	CBCT
2D	0.781^{**}_{**}	×	0.381^{**}_{*}	×	0.996^{*}_{*}	×	3 0.0	×
3D seg only	0.862	0.758_{**}	0.188	0.404	0.688	0.948	10.0	19.4
3D def only	0.861	0.797^{**}	0.180	0.256	0.672	0.787	0.00	0.00
3D joint - seg arm	0.870^{*}_{**}	0.783^{*}	0.169^{*}_{**}	0.251	0.671^{*}_{*}	0.808	0.00	0.00
3D joint - def arm	0.870^{*}_{*}	0.775^{*}_{*}	0.162^{*}_{**}	0.262	$0. 648^{*}_{**}$	0.807	0.00	0.00

CBCT segmentation. We thus trained and evaluated the 2D model on CT data only. The 3D segmentation-only model performed reasonably in CT segmentation, but a clear drop in performance is observed for CBCT segmentation. The deformation-only model performs comparable for CT segmentation, but considerably outperforms the segmentation-only model for CBCT segmentation. In comparison, the 3D joint model shows a synergistic effect for CT segmentation, outperforming both 3D baseline models. Yet, this synergy is not observed for CBCT segmentation. Note that the final predictions made in the two arms are not combined. However, after three refinement iterations, the predictions from both arms have converged to virtually the same segmentation.

3.3 | Visual evaluation

In Figure 6, sagittal views of the segmentation predictions are shown for the worst, median, and best performing segmentation instance of the 2D model. Segmentation predictions for both CT and CBCT are shown. The worst performing segmentation instance of the 2D model (top row) highlights the shortcomings of a 2D model. First, multiple connected components can be observed. Furthermore, as the model is prior-unaware, geometrical properties of the predicted segmentation are often inconsistent with the prior. The 3D prior-aware models on the other hand better propagate intrinsic shape characteristics from the prior segmentation. In the 3D segmentation only model, this property is not guaranteed, however. This becomes clear when looking at the CBCT segmentation results. For each of the shown cases, the segmentation predicted by the segmentation only model is not quite able to maintain volume and overall shape characteristics of the prior shape. In comparison, the deformation only and the joint models yield more accurate and more robust segmentation performance, especially for sub-optimal data such as CBCT. Between the two approaches, the segmentation cases shown below do not reveal clear differences.

4 | DISCUSSION

In this study, we compared different 3D segmentation network architectures that incorporate a shape prior. In

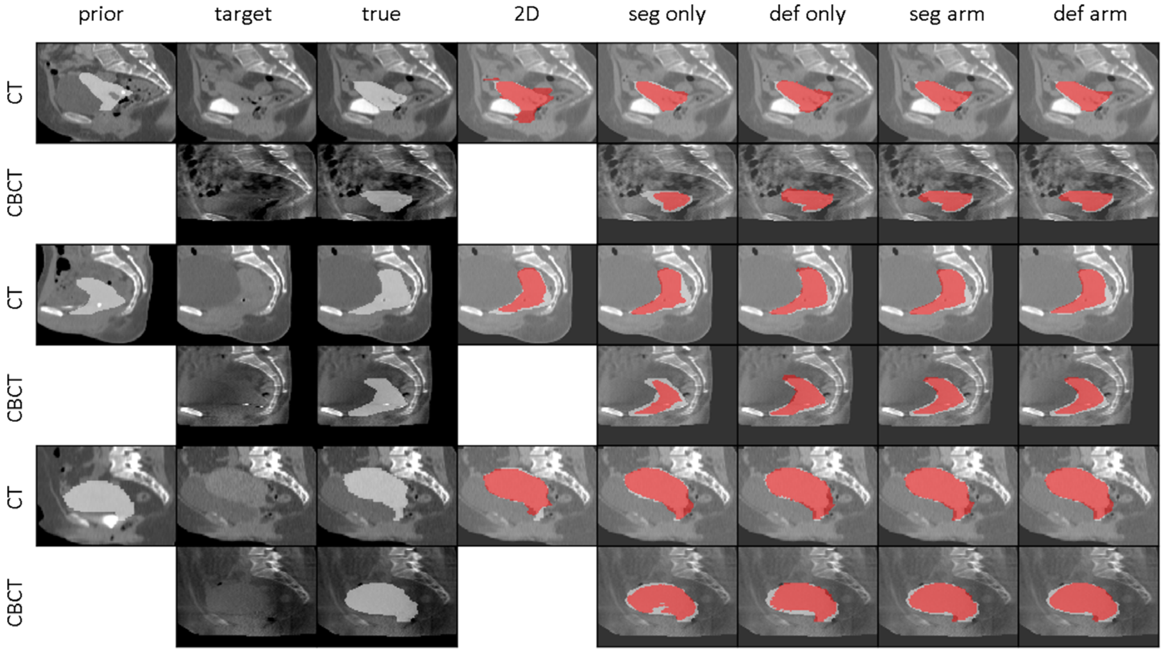

FIGURE 6 Segmentation predictions for both CT and CBCT for the worst (top), median (middle), and best (bottom) performing cases. Consecutively, the prior scan and segmentation, the target scan, the ground truth segmentations, and the segmentations predicted by the different models as indicated are shown. Predicted segmentations are shown in red, ground truth segmentations in white

particular, we proposed a new joint registration and segmentation model, which iteratively combines deformations of a provided prior with voxel-wise classification predictions. For reference, two additional prior-aware 3D segmentation networks were trained: a segmentation-only model, in which the prior is implicit by adding it to the network's input; and a deformation-only model, which iteratively optimizes a diffeomorphic deformation explicitly deforming the given prior. The 3D models were benchmarked against an established 2D segmentation network[8] that was tuned and optimized for this specific task.

We trained these models on the task of cervix-uterus CTV segmentation on both repeat CT scans as well as CBCT scans. We showed that, for our setting, regulating the segmentation prediction by the explicit propagation of a prior is beneficial as it improves the accuracy and robustness of the model's predictions. In common 2D or 3D segmentation models, segmentations are predicted as binary classification maps that assign independently to every pixel or voxel a probability of belonging to the segmentation. In comparison, by deforming a given prior segmentation, spatial dependency is explicit and is consistent with the prior by construction, see Figures 5 and 6. It is in this sense that the segmentation predictions are regulated.

Our setting benefits from this regularization as the segmentation model is challenged on three aspects. First, we used a relatively small patient cohort ($n = 84$) to train and evaluate the segmentation model on. Second, identifying soft-tissue contrast in the pelvic region can be difficult on CT scans and is further challenged by the inferior image quality of CBCT. Third, the CTV definition may be patient-specific, based on clinical considerations. While a common segmentation model will be unable to reproduce these clinical decisions, a model that deforms the planning segmentation may be able to propagate such decisions. Robustness to these conditions is important as they are not unique to our case, but commonly encountered in segmentation problems in the medical domain.

On CT scans, the segmentation-only and deformation-only models performed comparable, both substantially outperforming the 2D segmentation model. Some synergy was observed by combining both approaches in a joint network for CT, resulting in a median Dice score of about 0.87 and a median MSE surface distance below 1.7 mm (Table 2). Errors of these magnitude are comparable to manual delineation uncertainties. In Eminowicz et al.,[6] variability of manual delineation of the cervix-uterus CTV are reported to range between 0.72 and 0.90, expressed in terms of Dice coefficients. Moreover, these numbers suggest improved segmentation performance compared to recently published results on automatic CTV segmentation for cervical cancer,[35] in which a Dice score of 0.82 is reported.

On CBCT no clear synergistic effect was observed by combining both approaches. This is perhaps explained by considering the performance gap between the segmentation-only and deformation-only models for CBCT. Apparently, the network struggles to learn segmentations as per-voxel classification maps in challenging data settings such as CBCT. Consequently, incorporating a segmentation module into the network architecture may not be beneficial in such settings. Statistical power was also lower than for CT as differences occur less consistently in the same direction. Therefore, stating a clear winner is difficult, yet the models in which the prior segmentation is explicitly deformed clearly outperformed the segmentation-only model. This is in agreement with our hypothesis that the induced regularization by the deformation of a prior is especially important when the scan quality deteriorates. The deformation-only model performed best on our test set, with a median dice score of close to 0.8, and a median MSE surface distance of about 2.5 mm (Table 2). Addressing the clinical acceptability of these results requires further investigation and presumably depends on the application.

Automatic CBCT segmentation could be used in adaptive radiotherapy for automatic plan selection in a library of plans context or, more ambitiously, for daily replanning. The latter application would require accurate conversion from CBCT to synthetic CT. Efforts in this direction have been made, for example by training generative adversarial networks (GANs) such as the CycleGAN presented by Liang et al..[36] It would be interesting to see whether such conversion techniques can be used to further boost automatic segmentation accuracy.

Our findings are largely in agreement with those of Lee et al.,[12] who investigate the incorporation of shape priors in 3D networks for coronary artery segmentation on CT. Priors were incorporated both in a classification network, as well as in a template transformer network they called TeTrIS. Aside from some architectural differences, these models essentially correspond to the segmentation-only and deformation-only networks used for reference here. Similar to our results, they find that both methods perform similarly on CT and outperform the baseline model in which no prior is given. Furthermore, they qualitatively show in a toy example that the deformation network is more resilient to corruption of the input images. Again, our results reiterate this and bring this finding to a real-world scenario; that of automatic CBCT segmentation.

The training of a neural network can be adversely affected by inconsistent ground truth data. Human performance on segmentation tasks generally suffers from quite substantial inter- and intraobserver variability. The image quality has an effect on this, with observer variability increasing as images become more ambiguous.[37] Considering the poor soft-tissue contrast on CBCT, it is plausible that some of the performance drops on CBCT are caused by inconsistent ground truth annotations.[38]

Because of memory constraints, scans were downsampled in the axial plane. Effectively, this is a form of scan quality degradation and might therefore introduce a bias toward our finding that segmentation regularization by deformation of a prior is beneficial. It is not inconceivable that without this preprocessing, the performance of the 2D and 3D segmentation-only models would improve somewhat in absolute and relative terms.

Finally, it should be noted that the iterative procedure found in the deformation-only and joint networks is considerably slower than the segmentation-only network as it involves multiple iterations through a U-Net architecture instead of only one (Table 1). Also, the additional deformation operations require extra computational and memory resources. This makes the proposed deformation regularized models less suitable for the prediction of segmentation in real-time. However, this obstacle is merely practical and could be overcome by improved hardware or network architectures.

5 | CONCLUSION

We introduced a neural network for prior-aware autosegmentation that jointly predicts a voxel-wise classification map as well as a registration of the prior to the target image. By combining both approaches in a single neural network, improved segmentation performance could be achieved. More generally, we showed improved segmentation performance by constraining the predictions made to those that are reachable by a diffeomorphic deformation of the segmentation prior. This regularization becomes especially important when segmenting medical scans of low image quality.

ACKNOWLEDGMENTS

This research was funded by an internal grant from the Netherlands Cancer Institute.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

1. Taylor A, Powell MEB. An assessment of interfractional uterine and cervical motion: implications for radiotherapy target volume definition in gynaecological cancer. *Radiother Oncol*. 2008;88(2):250-257.
2. Bondar ML, Hoogeman MS, Mens JW, et al. Individualized nonadaptive and online-adaptive intensity-modulated radiotherapy treatment strategies for cervical cancer patients based on pretreatment acquired variable bladder filling computed tomography scans. *Int J Radiat Oncol Biol Phys*. 2012;83(5):1617-1623.
3. Beekman C, van Beek S, Stam J, Sonke JJ, Remeijer P. A biomechanical finite element model to generate a library of cervix CTVs. *Med Phys*. 2020;47(9):3852-3860.

4. Brock KK. Adaptive radiotherapy: moving into the future. *Semin Radiat Oncol*. 2019;29(3):181-184.
5. Dodge S, Karam L, Understanding how image quality affects deep neural networks. 2016 8th Int Conf Qual Multimed Exp QoMEX 2016. Published online 2016. https://doi.org/10.1109/QoMEX.2016.7498955
6. Eminowicz G, Mccormack M. Variability of clinical target volume delineation for definitive radiotherapy in cervix cancer. *Radiother Oncol*. 2015;117(3):542-547.
7. Lim K, Small W, Portelance L, et al. Consensus guidelines for delineation of clinical target volume for intensity-modulated pelvic radiotherapy for the definitive treatment of cervix cancer. *Int J Radiat Oncol Biol Phys*. 2011;79(2):348-355.
8. Nikolov S, Blackwell S, Zverovitch A, et al. Clinically applicable segmentation of head and neck anatomy for radiotherapy: deep learning algorithm development and validation study. *J Med Internet Res*. 2021;23(7).
9. Zheng S, Jayasumana S, Romera-Paredes B, et al. Conditional random fields as recurrent neural networks. Proceeding of IEEE International Conference on Computer Vision, Santiago, Chile, 7-13 December 2015. https://doi.org/10.1109/ICCV.2015.179
10. Dou Q, Yu L, Chen H, et al. 3D deeply supervised network for automated segmentation of volumetric medical images. *Med Image Anal*. 2017;41:40-54.
11. Dalca AV, Guttag J, Sabuncu MR. Anatomical priors in convolutional networks for unsupervised biomedical segmentation. *Proceeding of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 18-23 June 2018.
12. Lee MCH, Petersen K, Pawlowski N, Glocker B, Schaap M. TeTrIS: template transformer networks for image segmentation with shape priors. *IEEE Trans Med Imaging*. 2019;38(11):2596-2606.
13. Iglesias JE, Sabuncu MR. Multi-atlas segmentation of biomedical images: a survey. *Med Image Anal*. 2015;24(1):205-219.
14. Minaee S, Boykov YY, Porikli F, Plaza AJ, Kehtarnavaz N, Terzopoulos D. Image segmentation using deep learning: a survey. *IEEE Trans Pattern Anal Mach Intell*. 2021:1-22. Published online.
15. Mok TCW, Chung ACS. Large deformation diffeomorphic image registration with laplacian pyramid networks. *Lect Notes Comput Sci*. 2020;12263:211-221.
16. Ruder S. An overview of multi-task learning in deep neural networks. *ArXiv*. 2017. http://arxiv.org/abs/1706.05098. Published online.
17. Riklin-Raviv T, Kiryati N, Sochen N. Prior-based segmentation and shape registration in the presence of perspective distortion. *Int J Comput Vis*. 2007;72(3):309-328.
18. Yeo D, Lee CO. Variational shape prior segmentation with an initial curve based on image registration technique. *Image Vis Comput*. 2020;94:103865.
19. Chen X, Brady M, Rueckert D. Simultaneous segmentation and registration for medical image. *Lect Notes Comput Sci*. 2004;3216:663-670.
20. Pohl KM, Fisher J, Grimson WEL, Kikinis R, Wells WM. A Bayesian model for joint segmentation and registration. *Neuroimage*. 2006;31(1):228-239.
21. Zhu W, Myronenko A, Xu Z, et al. NeurReg: neural registration and its application to image segmentation. *Proceeding of 2020 IEEE Winter Conference Application of Computer Vision WACV*, Snowmass, CO, USA, 1-5 March 2020.
22. Estienne T, Vakalopoulou M, Christodoulidis S, et al. U-ReSNet: ultimate coupling of registration and segmentation with deep nets. *Lect Notes Comput Sci*. 2019;11766:310-319.
23. Xu Z, Niethammer M. DeepAtlas: Joint semi-supervised learning of image registration and segmentation. *Lect Notes Comput Sci*. 2019;11765:420-429.
24. Zhao S, Dong Y, Chang E, Xu Y. Recursive cascaded networks for unsupervised medical image registration. Proceeding of IEEE International Conference Computer Vision, Seoul, Korea (South), 27 October–2 November 2019.
25. Zhao S, Lau T, Luo J, Chang EIC, Xu Y. Unsupervised 3D end-to-end medical image registration with volume tweening network. *IEEE J Biomed Heal Informatics*. 2020;24(5):1394-1404.
26. Beekman C, Schaake E, Sonke J-J, Remeijer P. Deformation trajectory prediction using a neural network trained on finite element data—application to library of CTVs creation for cervical cancer. *Phys Med Biol*. 2021;66(21):215004.
27. Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-net: learning dense volumetric segmentation from sparse annotation. *Lect Notes Comput Sci*. 2016;9901:424-432.
28. Balakrishnan G, Zhao A, Sabuncu MR, Guttag J, Voxel-Morph DalcaAV. A learning framework for deformable medical image registration. *IEEE Trans Med Imaging*. 2019;38(8):1788-1800.
29. Cui Y, Jia M, Lin TY, Song Y, Belongie S. Class-balanced loss based on effective number of samples. Proceeding of IEEE Computer Society Conference Computer Vision Pattern Recognition, Long Beach, CA, USA, 15-20 June 2019.
30. Ashburner J. A fast diffeomorphic image registration algorithm. *Neuroimage*. 2007;38(1):95-113.
31. Arsigny V, Commowick O, Pennec X, Ayache N. A log-euclidean framework for statistics on diffeomorphisms. *Lect Notes Comput Sci*. 2006;4190:924-931.
32. Dalca AV, Balakrishnan G, Guttag J, Sabuncu MR. Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces. *Med Image Anal*. 2019;57:226-236.
33. Dou Q, Liu Q, Heng PA, Glocker B. Unpaired multi-modal segmentation via knowledge distillation. *IEEE Trans Med Imaging*. 2020;39(7):2415-2425.
34. Izmailov P, Podoprikhin D, Garipov T, Vetrov D, Wilson AG. Averaging weights leads to wider optima and better generalization. *34th Conf Uncertain Artif Intell 2018 UAI 2018*. 2018;2:876-885.
35. Ju Z, Guo W, Gu S, et al. CT based automatic clinical target volume delineation using a dense-fully connected convolution network for cervical Cancer radiation therapy. *BMC Cancer*. 2021;21(1):1-10.
36. Liang X, Chen L, Nguyen D, et al. Generating synthesized computed tomography (CT) from cone-beam computed tomography (CBCT) using CycleGAN for adaptive radiation therapy. *Phys Med Biol*. 2019;64(12):125002.
37. Lütgendorf-Caucig C, Fotina I, Stock M, Pötter R, Goldner G, Georg D. Feasibility of CBCT-based target and normal structure delineation in prostate cancer radiotherapy: multi-observer and image multi-modality study. *Radiother Oncol*. 2011;98(2):154-161.
38. Karimi D, Dou H, Warfield SK, Gholipour A. Deep learning with noisy labels: exploring techniques and remedies in medical image analysis. *Med Image Anal*. 2020;65:1-22.

How to cite this article: Beekman C, van Beek S, Stam J, Sonke J-J, Remeijer P. Improving predictive CTV segmentation on CT and CBCT for cervical cancer by diffeomorphic registration of a prior. *Med Phys.* 2022;49:1701–1711. https://doi.org/10.1002/mp.15421