Deep learning-based segmentation in prostate radiation therapy using Monte Carlo simulated cone-beam CT

Nelly Abbani, Thomas Baudier, Simon Rit, Francesca Di Franco, Franklin Okoli, Vincent Jaouen, Florian Tilquin, Anaïs Barateau, Antoine Simon, Renaud de Crevoisier, et al.

Deep learning-based segmentation in prostate radiation therapy using Monte Carlo simulated cone-beam CT

Ruuning Title: Deep learning CBCT segmentation

Nelly Abbani*,[1] Thomas Baudier,[1] Simon Rit,[1] Francesca di Franco,[1] Franklin Okoli,[2] Vincent Jaouen,[2] Florian Tilquin,[3] Anaïs Barateau,[3] Antoine Simon,[3] Renaud de Crevoisier,[3] Julien Bert,[2] and David Sarrut[1]

[1]Univ Lyon, INSA-Lyon, Université Claude Bernard Lyon 1, CNRS, Inserm, CREATIS UMR 5220, U1206, F-69621, LYON, France.

[2]LaTIM, Université de Bretagne Occidentale, INSERM UMR 1101, IMT Atlantique, CHRU Brest, Brest, France.

[3]Univ Rennes, CLCC Eugène Marquis, Inserm, LTSI – UMR 1099, F-35000 Rennes, France

(Dated: August 20, 2022)

* Author to whom correspondence should be addressed
Email: nelly.abbani@gmail.com

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/mp.15946

Abstract

Purpose. Segmenting organs in cone-beam CT (CBCT) images would allow to adapt the radiotherapy based on the organ deformations that may occur between treatment fractions. However, this is a difficult task because of the relative lack of contrast in CBCT images, leading to high inter-observer variability. Deformable image registration (DIR) and deep-learning based automatic segmentation approaches have shown interesting results for this task in the past years. However, they are either sensitive to large organ deformations, or require to train a convolutional neural network (CNN) from a database of delineated CBCT images, which is difficult to do without improvement of image quality. In this work, we propose an alternative approach: to train a CNN (using a deep learning-based segmentation tool called nnU-Net) from a database of artificial CBCT images simulated from planning CT, for which it is easier to obtain the organ contours.

Methods. Pseudo-CBCT (pCBCT) images were simulated from readily available segmented planning CT images, using the GATE Monte Carlo simulation. CT reference delineations were copied onto the pCBCT, resulting in a database of segmented images used to train the neural network. The studied segmentation contours were: bladder, rectum, and prostate contours. We trained multiple nnU-Net models using different training: 1) segmented real CBCT, 2) pCBCT, 3) segmented real CT and tested on pseudo-CT (pCT) generated from CBCT with cycleGAN, and 4) a combination of 2) and 3). The evaluation was performed on different datasets of segmented CBCT or pCT by comparing predicted segmentations with reference ones thanks to Dice similarity score and Hausdorff distance. A qualitative evaluation was also performed to compare DIR-based and nnU-Net-based segmentations.

Results. Training with pCBCT was found to lead to comparable results to using real CBCT images. When evaluated on CBCT obtained from the same hospital as the CT images used in the simulation of the pCBCT, the model trained with pCBCT scored mean DSCs of 0.92 ± 0.05, 0.87 ± 0.02, and 0.85 ± 0.04 and mean Hausdorff distance 4.67 ± 3.01, 3.91 ± 0.98, and 5.00 ± 1.32 for the bladder, rectum, and prostate contours respectively, while the model trained with real CBCT scored mean DSCs of 0.91 ± 0.06, 0.83 ± 0.07, and 0.81 ± 0.05 and mean Hausdorff distance 5.62 ± 3.24, 6.43 ± 5.11, and 6.19 ± 1.14 for the bladder, rectum, and prostate contours respectively. It was also found to outperform models using pCT or a combination of both, except for the prostate contour when tested on a dataset from a different hospital. Moreover, the resulting segmentations demonstrated a clinical acceptability, where 78% of bladder segmentations, 98% of

rectum segmentations, and 93% of prostate segmentations required minor or no corrections, and for 76% of the patients, all structures of the patient required minor or no corrections.

Conclusion. We proposed to use simulated CBCT images to train a nnU-Net segmentation model, avoiding the need to gather complex and time-consuming reference delineations on CBCT images.

Keywords: deep learning, segmentation, CBCT, Monte Carlo simulation, prostate, cancer

I. INTRODUCTION

In adaptive radiotherapy of pelvic treatments, the delineation of organs at risk (OAR) from cone-beam computed tomography (CBCT) images is an important step to ensure proper OAR sparing, because anatomical deformations may occur between the treatment fractions and may not be accounted for in the treatment plan, leading to uncertainties in the dose distribution[1]. However, this is a difficult and time consuming task that is generally not routinely done in clinical practice because of the relative lack of contrast in CBCT images, leading to high inter-observer variability[2].

One common strategy proposed for CBCT segmentation is to perform deformable image registration (DIR) between the treatment planning CT and the CBCT images in order to deform contours delineated on CT images to CBCT images[3]. Another method is to use multi-atlas based segmentation approaches for propagating manual delineations to CBCT images[4]. However, DIR is sensitive to large organ deformations (e.g. bladder), which may lead to unacceptable segmentation accuracy[5,6]. Another limitation is the amplification of segmentation errors through contour deformation.

Recently, deep learning (DL) based automatic segmentation approaches have shown impressive results compared to conventional methods[7], especially for bladder, prostate and rectum segmentation in CT images[8]. However, such methods may be harder to apply to CBCT images due to the lower quality of these images compared to CT, with poor soft tissue contrast and high noise levels. Moreover, those supervised methods require the training of a convolutional neural network (CNN) from a database of segmented CBCT images, which can be difficult to obtain and may contain relatively large contours uncertainties[9] compared to contours made on CT[2].

There have been multiple attempts to enhance the quality of CBCT images. One approach is to perform a scatter correction of the CBCT projections, e.g. using scatter kernel algorithms[10], a CT image of the same patient[11] or an anti-scatter grid[12]. Another approach is to train a generative deep learning network to synthesize CT images[13–16]. The training database is then made of delineated planning CT images, which can be obtained from existing clinical databases. The CBCT image is first transformed into a pseudo-CT (pCT, for example with cycleGAN) before being processed by the CNN for segmentation. Using synthesized images from GAN is still in its infancy and the limitations of it are still unclear.

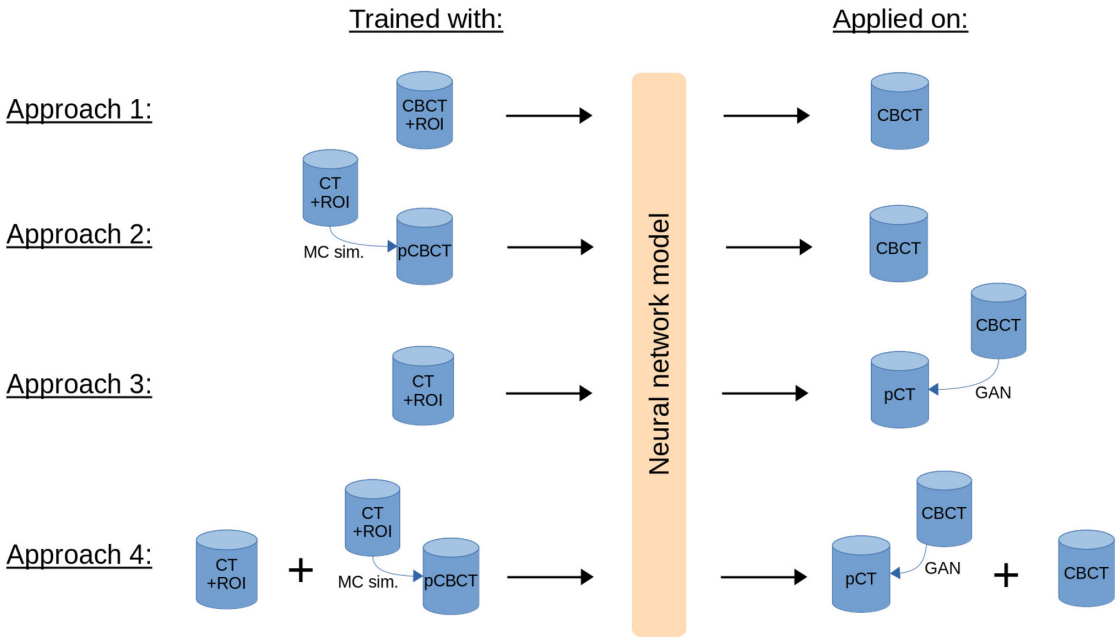

FIG. 1: Different trained neural network models for auto CBCT image segmentation, using 1) segmented CBCT images, 2) pCBCT images simulated from CT, 3) CT images and pCT generated from CBCT, and 4) a combination of the previous two.

In the continuity of those ideas, we propose to explore an original auto segmentation approach. Instead of training a CNN from segmented planning CT images and using pseudo-CT generated from CBCT, we propose to train the CNN from pseudo-CBCT images simulated from planning CT using physically realistic Monte Carlo simulations. We compare CNN trained 1) from segmented CBCT images, 2) from pCBCT images, 3) from CT images + using pCT, and 4) a combination of 2) and 3), as summarized in figure 1. Note that the neural network used in this work is a deep learning-based segmentation tool called no-new U-Net (nnU-Net), however, we refer here to CNN in general because alternative neural networks could be used.

II. MATERIALS AND METHODS

A. Database of CT and CBCT images

The first database DB_{CT} contains data from 90 prostate cancer patients, collected from the radiotherapy department of the Léon Bérard cancer center (CLB, Lyon, France). A total of 90 CT was available, acquired between December 2011 and November 2019. All data were anonymized and respected GDPR and local regulations for patient privacy. For each patient, the database includes: 1) the planning CT image, 2) the contours of pelvic organs, manually delineated by experts on the planning CT during the treatment planning, 3) several CBCT images, 4) geometrical information for each CBCT acquisition, such as the gantry angles and detector offsets at each projection. The CBCT images in DB_{CT} have not been used in this study, only the planning CT of each patient and the geometrical information for the simulation of CBCT images from the planning CT (see section II C). The size of the CT images varies from patient to patient, but it is usually 512 x 512 pixels per slice, with a pixel size of about 1 mm in axial slices, and 3 mm in the axial direction. CT images have been acquired with a Philips Gemini Big Bore CT system or with a Siemens Confidence 20 scanner using 120 kVp and a tube current of 146 mA. CT contours have been delineated once by different physicians and medical physicists in charge of the patient treatments. Several contours of volumes of interest (VOI) were available: bladder, prostate, rectum, femoral heads, and seminal vesicles. Note that rectums were generally only delineated in the region of irradiation and not entirely (usually 2 cm below and above the prostate). All contours have been converted to 3D binary mask images. CBCT images have a size of $410 \times 410 \times 264$ voxels and spacing $1 \times 1 \times 1$ mm. They were acquired with an Elekta Synergy XVI device and a bow-tie filter, a voltage of 120 kV, a tube current of 40 mA, and an exposure time of 40 ms/projection, which are the presets suggested by the vendor for imaging the pelvis.

Two other databases $DB_{CBCT,\ CLB}$ and $DB_{CBCT,\ CEM}$ were built with CBCT images from patients different from those in DB_{CT}, where several VOI have been delineated on the CBCT images by experts. The contours for the images in $DB_{CBCT,\ CEM}$ were manually delineated by multiple physicians, while those for $DB_{CBCT,\ CLB}$ were first obtained using a DIR auto-segmentation tool (ADMIRE v3.26, Elekta AB, Stockholm, Sweden), and then reviewed and corrected by a physician. 41 images (of size $410 \times 410 \times 264$ voxels and $1 \times 1 \times 1$ mm

spacing), acquired between November 2020 and June 2021, were collected from 9 patients at the Léon Bérard cancer center (CLB, Lyon, France), and 130 images of size $410 \times 410 \times 168$ voxels and $1 \times 1 \times 1$ mm spacing, acquired between February 2010 and April 2014, were collected from 6 patients at the Eugène Marquis cancer center (CEM, Rennes, France). All images were acquired with an Elekta Synergy XVI. Bladder, prostate, and rectum contours were considered here. Those two databases were used to evaluate the accuracy of the auto-segmentation on images collected from different sources. Indeed, because CT contours in DB_{CT} and reference CBCT contours in $DB_{CBCT,\,CLB}$ have been performed in the same hospital (CLB), the second reference CBCT dataset $DB_{CBCT,\,CEM}$ with contours performed by different physicians may contain differences due to the inter-observer variability, and hence provides an additional challenge to the proposed model. Obtaining a dataset of segmented CBCT is a tedious task because CBCT images generally have a low contrast, especially in the pelvic region. Here, while the number of patients included is limited (9+6), the number of images is rather large (41+130), and is comparable to other similar studies in that field (between 6 and 15 patients, 15 and 115 images[17–20]). Table I summarizes the database properties.

TABLE I: Available databases

Database	**Num. Images**	**Size**	**Machine Type**	**Available VOI**
DB_{CT} (CLB)	90 CT (90 patients)	512×512/slice approx. $1 \times 1 \times 3$ [mm/px]	Philips Gemini Big Bore or Siemens Confidence 20	Bladder, prostate, rectum, femoral heads, and seminal vesicles
$DB_{CBCT,\,CLB}$ (CLB)	41 CBCT (9 patients)	$410 \times 410 \times 264$ px $1 \times 1 \times 1$ [mm/px]	Elekta Synergy XVI	Bladder, prostate, rectum
$DB_{CBCT,\,CEM}$ (CEM)	130 CBCT (6 patients)	$410 \times 410 \times 168$ px $1 \times 1 \times 1$ [mm/px]	Elekta Synergy XVI	Bladder, prostate, rectum

B. Deformable image registration

DIR estimates the geometric transformation that warps one image in order to maximise the similarity to another image, which may be from a different modality. It is a conventional technique used in adaptive treatment radiotherapy as it deals with organ deformation between images, by warping not only the image, but other information attached to the image, such as anatomical contours and radiation dose[21]. Contours for the 41 CBCT images in $DB_{CBCT,\ CLB}$ were obtained using a DIR auto-segmentation tool, Advanced Medical Imaging Registration Engine (ADMIRE) v3.26 (Elekta AB, Stockholm, Sweden), which uses the image correspondence between the planning CT and the CBCT in order to propagate the CT contours to the CBCT[3]. These contours will be used to compare the proposed CNN-based auto-segmentation method to a DIR-based method commercially developed and currently used in the clinic.

C. Pseudo-CBCT generation from CT via simulation

We propose to simulate pseudo-CBCT (pCBCT) images from CT images to exploit the associated contours that are delineated by clinicians for treatment planning. Once the pCBCT is created, all CT contours are directly aligned with the pCBCT and can be used to train a network for auto-contouring from real CBCT images. We used the 90 CT images in DB_{CT} to built DB_{pCBCT} , a database of 90 simulated pCBCT that will be used in training some of our models.

pCBCT images were created by reconstruction from radiographic projections generated from the planning CT image with a Monte Carlo simulation of the CBCT scanner using the simulation software GATE[22,23]. The input planning CT image was positioned and oriented in the simulation according to the treatment machine log files available for each patient in the initial database DB_{CT}. The source and the detector of the Elekta XVI CBCT scanner were simulated using estimated source spectrum and detector response from measurements[24] and geometrical information (position and orientation) available in DB_{CT}. The output x-ray projections at each angle were simulated using Fixed Forced Detection (FFD)[25,26], a variance reduction technique which simulates the scatter by mixing Monte Carlo and deterministic simulation to accelerate conventional Monte Carlo simulations. The simulation of

the primary radiation is completely deterministic, i.e., it computes the source-to-detector line integrals of the attenuation coefficients with ray tracing. Each projection was divided by the flat field image, i.e., the projection simulated without CT image. The simulated projections were reconstructed using RTK[27], a CBCT reconstruction toolkit based on ITK. The projections had a size of 512 × 512 pixels and 0.8 × 0.8 mm pixel size, and the output image had a size of 410 × 410 × 264 voxels and 1 mm isotropic voxel size, similar to that of the real projections and CBCT images of DB_{CT}. After reconstruction, voxels outside the field of view were set to 0. The reconstructed voxel values represent the photon attenuation coefficient μ, which was converted to CBCT numbers with $CBCT\# = \mu \times 2^{16} - 1024$ to mimick the scanner processing indicated in previous works[26,28].

D. Pseudo-CT generation from CBCT via cycleGAN

Another strategy would be to use a network trained with delineated planning CT to segment pseudo-CT images generated from CBCT images. The labels predicted by the network could then be transferred to the corresponding CBCT image. This is for example done by Zhao et al.[15] To that end, two databases of pseudo-CT, $DB_{pCT,\ CLB}$ (41 images) and $DB_{pCT,\ CEM}$ (130 images) were generated from $DB_{CBCT,\ CLB}$ and $DB_{CBCT,\ CEM}$ respectively using a cycleGAN[29,30]. The delineations of each CBCT were transferred to the corresponding pCT, to evaluate the result of its auto-segmentation by the CNN network. The cycleGAN is based on the combination of two generative adversarial networks (GANs) working in parallel. One of them was trained to provide a mapping from the CBCT image space to the CT image space, and the other one conversely. Both GAN were trained together, alternating back-propagation during each iteration. A loss function was computed over the composition of both mappings, enforcing that coming back to the original image space provided a coherent intensity distribution. It was combined with the loss functions of the GANs, based on a least-square objective function for both the generator and discriminator. The cycleGAN was trained on CT and CBCT images of 18 patients from the CEM hospital treated for prostate cancer (completely different from the patients included in $DB_{CBCT,\ CEM}$). The number of images, while limited, is close to the ones reported in other studies generating pCT from CBCT (between 5 and 205[31]).

E. Deep learning-based image segmentation

We used *no-new U-Net* (nnU-Net), a self-configuring deep learning-based segmentation method[32] to investigate automated segmentation from pCBCT and pCT. nnU-Net is characterized by its ability to self adapt to new datasets, by automatically adjusting the U-Net model parameters based on the properties of the training dataset. For the training images from DB_{CT} and $DB_{CBCT,\,CLB}$, the patch size set by nnU-Net was $191 \times 257 \times 219$ voxels, with a batch size of 2, and minimum feature map size of 32. For the images from $DB_{CBCT,\,CEM}$, the patch size was $178 \times 308 \times 233$ voxels. In nnU-Net, each image is normalized independently using z-scoring, except for CT images where a global normalization scheme is applied on the whole dataset by clipping to the [0.5, 99.5] percentiles of all intensity values, then applying a z-score normalization based on the mean and standard deviation of all collected values. This is done because intensity values in CT images are quantitative and reflect physical properties of the tissue, and so it can be beneficial to retain this information[32]. Since the images used in this project are CT and CBCT images, so with quantitative intensity values, the global normalization scheme was applied. In addition, nnU-Net automatically applies, stochastically according to a predefined probability, a variety of data augmentation techniques during training: rotations, scaling, gamma correction and mirroring[33].

Other model parameters are configured automatically by the built-in nnU-Net trainers regardless of the provided dataset. The default architectures use plain convolutions, instance normalization and Leaky ReLU activation function. Downsampling is done with strided convolutions, upsampling is done with convolutions transposed, and two computational blocks are used per resolution stage both in the encoder and the decoder[32]. The default optimizer is a stochastic gradient descent with a high initial learning rate (0.01) and a large Nesterov momentum (0.99). The loss function is the sum of cross-entropy and Dice loss. nnU-Net also uses five-fold cross validation training, where the training dataset is automatically divided into five folds to train five configurations, each using one subset as a validation dataset. This allows to use the entire training set for validation, improving the accuracy of the inference which uses the ensemble of these five configurations to predict the test cases.

In this study, the number of epochs was set to 200 instead of the default 1000 in order to save computation time while still allowing convergence of the loss function.

F. Trained models

1. Datasets

As previously mentioned, we investigated the interest of using pCBCT images for training nnU-Net compared to using real CT images or real CBCT images. Table II summarizes the datasets used for training the different models.

One nnU-Net (M3) was trained using the pCBCT images simulated from DB_{CT} (approach 2 in figure 1), another nnU-Net (M4) using the real CT images, of the same patients, from DB_{CT} (approach 3 in figure 1), and a third one (M5) using both pCBCT and CT as two channels multi-modal input images (approach 4 in figure 1). Since the pCBCT have been simulated from the planning CT, the training datasets M3, M4, and M5 share the same reference labels. Similarly, the reference labels for the CBCT (from $DB_{CBCT,\ CLB}$ and $DB_{CBCT,\ CEM}$) and pCT (from $DB_{pCT,\ CLB}$ and $DB_{pCT,\ CEM}$) of the test datasets of these models are the same. Additionally, two reference models (M1 and M2) were trained using real delineated CBCT images (approach 1 in figure 1), once using those obtained from CEM, and once using those from CLB. Note that the validation dataset is automatically defined and used by the five-fold process of nnU-Net (see section II E).

TABLE II: Datasets used for the training of the 5 models.

Model	Training datasets	Test datasets
M1	$\mathbf{DB_{CBCT,\ CLB}}$ (41 real CBCT from CLB)	$\mathbf{DB_{CBCT,\ CEM}}$ (130 real CBCT from CEM)
M2	$\mathbf{DB_{CBCT,\ CEM}}$ (130 real CBCT from CEM)	$\mathbf{DB_{CBCT,\ CLB}}$ (41 real CBCT from CLB)
M3	$\mathbf{DB_{pCBCT}}$ (90 pCBCT)	$\mathbf{DB_{CBCT,\ CEM}}$ (130 real CBCT from CEM) $\mathbf{DB_{CBCT,\ CLB}}$ (41 real CBCT from CLB)
M4	$\mathbf{DB_{CT}}$ (90 real CT)	$\mathbf{DB_{pCT,\ CEM}}$ (130 pCT from CEM) $\mathbf{DB_{pCT,\ CLB}}$ (41 pCT from CLB)
M5	$\mathbf{DB_{CT}}$ + $\mathbf{DB_{pCBCT}}$ (90 real CT + pCBCT)	$\mathbf{DB_{pCT,\ CEM}}$ + $\mathbf{DB_{CBCT,\ CEM}}$ (130 pCT + real CBCT from CEM) $\mathbf{DB_{pCT,\ CLB}}$ + $\mathbf{DB_{CBCT,\ CLB}}$ (41 pCT + real CBCT from CLB)

2. Experimental settings

All training and test images were cropped and resampled to a size of 410 × 410 × 264 voxels with 2 mm spacing. The images in all training datasets were further cropped along the coronal plane to the area where at least a reference contour was available because the rectum was, in general, not delineated in all axial slices. It was expected that this would improve the performance of the model.

3. Evaluation metrics

The Dice similarity coefficient (DSC) was used to evaluate the accuracy of the contour prediction of the models. Since all test datasets have the same reference labels, it is possible to provide a direct comparison between the scores of each model on both the CEM and CLB datasets. It was calculated only with the slices where a reference delineation was available. For the rectum, it was only calculated in the slices were a prediction was available because the rectum was not delineated similarly between all hospitals (delineation length differs from hospital to hospital). We also computed the 95th percentile of the Hausdorff distance (computed with MedPy[34]). It measures how far the predicted contours are from the reference contour, by calculating the 95th percentile of longest surface distance from any point in one set to the closest point in the other set. Using the 95th percentile as opposed to the maximum distance is preferable to make the metric less sensitive to outliers[34]. Both metrics were evaluated for the 3 soft tissues contours: bladder, rectum, and prostate. To evaluate the statistical significance of the difference in the DSCs and Hausdorff distances of the different models, the Wilcoxon signed-rank test (from SciPy[35]) was used since the distribution of the differences between these sets cannot be assumed to be normally distributed.

G. DIR contours comparison

The contours generated by our method (M3) were also compared to DIR-based contours from ADMIRE (see section II B). However, it was not possible to directly compare DSC computed between M3 and reference $DB_{CBCT,\,CLB}$ on one hand, and between ADMIRE and reference $DB_{CBCT,\,CLB}$ on the other hand, because the reference contours were obtained in a two steps process, starting from the DIR auto contours from ADMIRE, followed by

manual refinement and correction by the clinician expert. Hence, the final reference contours are not independent of the DIR-based contours. Instead, we proceeded with the following qualitative blind process. An expert from CLB was provided with 82 randonmly ordered CBCT images, each image from $DB_{CBCT,\,CLB}$ occurring twice, once with contours obtained using nnU-Net (M3), and once with contours obtained using ADMIRE. Not knowing the source of the contours for each image, the clinician visually evaluated the contours (bladder, rectum, prostate) based on the grading scheme presented in table III, proposed by Schreier et al.[19] The Wilcoxon signed-rank test was performed to evaluate the statistical significance of the differences between the scores of the DIR- and nnU-Net-based contours. Moreover, Schreier et al.[19] defined two criteria to evaluate the clinical acceptability of the effort needed to correct automatically generated contours: 1) for each structure, more than 80% of the patients receive a score of 2 or 3 for that structure, and 2) for more than 70% of all patients, all structures of this patient receive a score of 2 or 3.

TABLE III: The grading scheme used for the qualitative evaluation.

Score	Definition
0	Not acceptable, manual (re)drawing of the entire structure is required
1	Acceptable, major corrections necessary but with acceptable effort. Corrections on more than 5 slices
2	Accepted, only minor corrections required. Corrections on less than 5 slices
3	Accepted, no corrections required

III. RESULTS

A. CBCT simulation

Figure 2 shows the real CT and real CBCT of a patient from DB_{CT}, and the pCBCT simulated from the CT for comparison, and the histogram and intensity profiles of the 3D images. On the image slices we can see that the pCBCT has the same field-of-view and similar quality as the real CBCT.

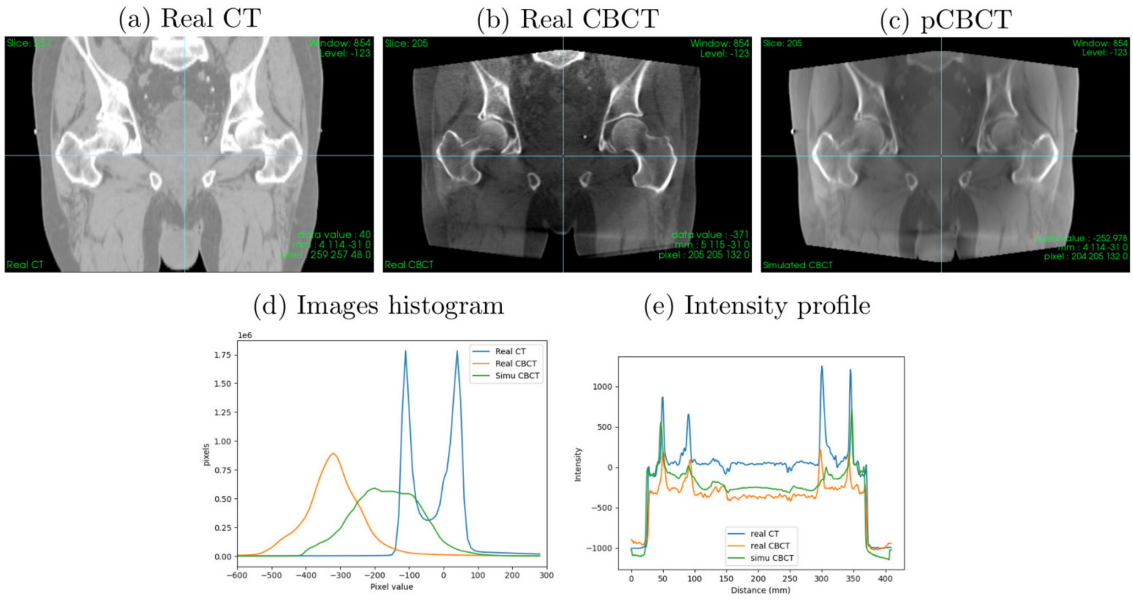

FIG. 2: Image slices of real CT, real CBCT and pCBCT (top), histogram of the 3D images (bottom left) and intensity profiles (bottom right) taken in the left-right direction (blue horizontal lines in the top images).

B. Pseudo CT generation

For evaluating the performance of the GAN, the mean absolute error (MAE) was calculated between the pseudo-CT generated by the GAN (86 pseudo-CT from 5 patients, completely independent from the patients in the training dataset), and the reference CT of the same patient. The MAE in the patient contour for this GAN was 38.4 HU, which is in accordance with MAE values reported by Spadea et al.[31] (between 16 and 87 HU) for pelvic pseudo-CT generation in other studies.

Figure 3 shows the real CT, real CBCT, and the pCT simulated from the CBCT of a patient from $DB_{CBCT,\,CEM}$. The histograms and intensity profiles images are also included. We notice that the quality and contrast in the pCT is similar to the real CT.

C. Loss plots

During a model training, nnU-Net provides a set of output files that monitor the progress of the training. One file includes the plot of the training and validation loss during training,

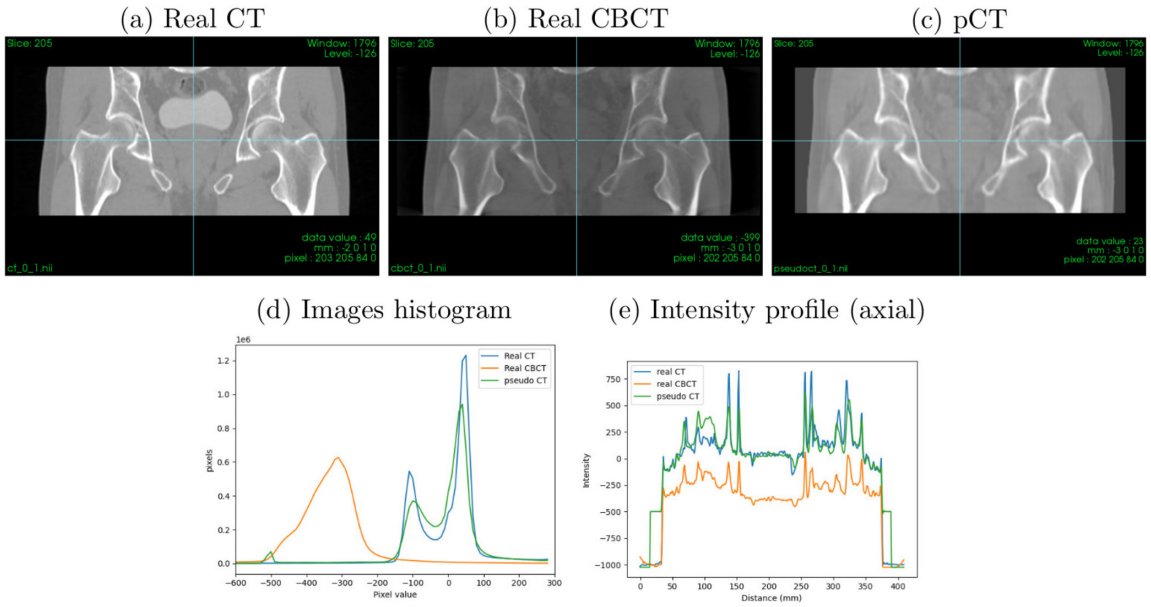

FIG. 3: Image slices of real CT, real CBCT and pCT (top), histogram of the 3D images (bottom left) and intensity profiles (bottom right) taken in the left-right direction (blue horizontal lines in the top images).

as well as an approximation of the evaluation metric, which is the average Dice score of the foreground classes, computed on randomly drawn patches from the validation data at the end of each epoch[33]. The Dice loss ranges from 0 to -1 and cross-entropy loss from infinity to 0, so the best loss is -1. Dice scores range from 0 to 1, so the best evaluation metric is 1. Figure 4 displays the plot for one of the trained models. As we can see, this model converges by the end of the 200 epochs.

D. CBCT segmentation visual evaluation

Figure 5 shows the reference labels (yellow), M3 prediction labels (red), and M4 prediction labels (green) on real CBCT images for 6 different patients from the CLB and CEM cohorts. As we can see in figure (b), the rectum is not entirely delineated in the reference label, and while the models correctly contoured a larger area of the rectum, the DSC and HD in these slices would have indicated a worse performance of the model. To prevent this, the slices where a prediction is available and a reference label is not provided are excluded from the DSC and HD calculation.

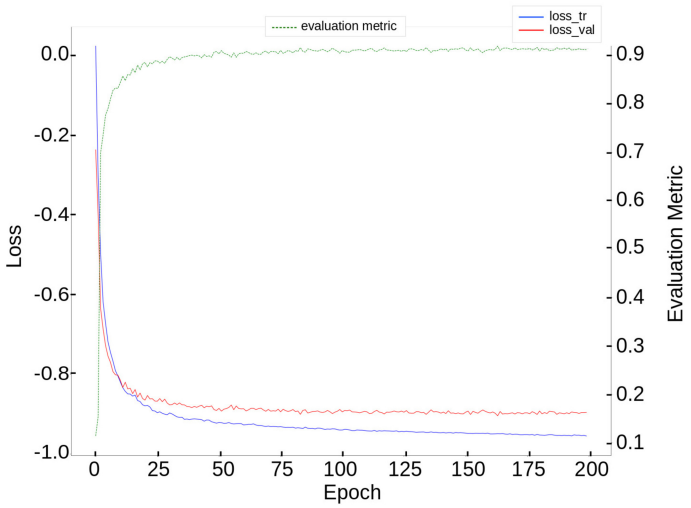

FIG. 4: Plot of the training (blue) and validation (red) loss and the evaluation metric (green) against epochs for one of the trained models.

E. CBCT segmentation evaluation metrics

Figure 6 displays the violin plot of the DSCs for the different models (summarized in table II), evaluated with both CLB and CEM test datasets. Figure 7 displays the violin plot of the Hausdorff distance from the same tests. The dashed red line allows to visualize the mean value of the models with respect to that for M3. A higher mean DSC or lower mean HD implies a better performance of that model in comparison to M3. The stars below or above each violin plot indicate the statistical significance of the difference between the results of each model with that of M3 (pCBCT), based on the p-value computed with the Wilcoxon signed-rank test. The "ns" symbol represents a p-value above 0.05 (no statistical difference), one star represents a p-value between 0.001 and 0.05 (statistically significant difference), and two stars represent a p-value below 0.001 (highly statistically significant difference). The numbers above the plots indicate the number of outliers for each violin plot, which consists of the values that are below $(Q1 - 1.5 \times IQR)$ or above $(Q3 + 1.5 \times IQR)$, where Q1 and Q3 are the first and third quantiles respectively, and the interquartile (IQR) range is IQR $= Q3 - Q1$. It provides a more comprehensive evaluation of the models as an elevated number of outliers indicates a low or inconsistent performance of the model with too many values falling outside the boundaries of the violin plot. The first violin plot (M1 or M2) on the left of all panels corresponds to the conventional training from real CBCT images,

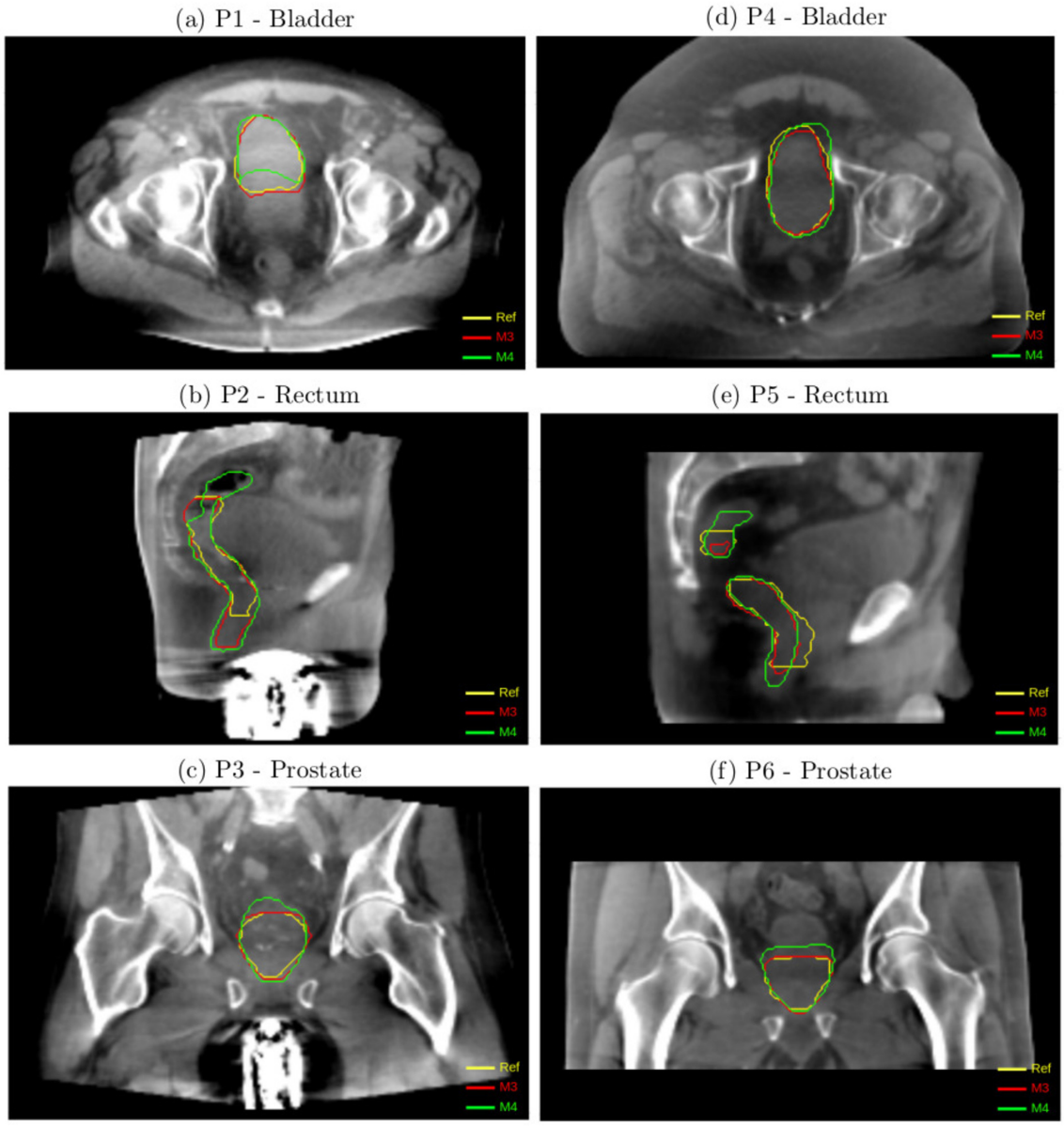

FIG. 5: Image slices of real CBCT of 6 different patients, 3 from CLB (P1 - P3) and 3 from CEM (P4 - P6), with reference labels (yellow), M3 prediction labels (red), and M4 prediction labels (green) for the bladder, rectum, and prostate labels.

evaluated on real CBCT images. The second violin plot (M3) corresponds to training with simulated pCBCT images, evaluated on real CBCT images. The third violin plot in the panels (M4) corresponds to training with real CT images, evaluated on pseudo-CT images. The last violin plot (M5) corresponds to combined training using real CT images in channel

0 and simulated pCBCT images in channel 1, evaluated on pseudo-CT in channel 0 and real CBCT in channel 1.

Table IV summarizes the mean DSC and Hausdorff distance for the different models, with the same semantic for the statistical significance tests.

TABLE IV: DSCs and Hausdorff distance mean values. The stars next to the values indicate the statistical significance of the difference between the scores for that model and that of M3 for the same metric and same structure, based on the p-value computed with the Wilcoxon signed-rank test.

Dataset	model	Mean DSC (Bladder)	Mean HD (mm) (Bladder)	Mean DSC (Rectum)	Mean HD (mm) (Rectum)	Mean DSC (Prostate)	Mean HD (mm) (Prostate)
CLB	M2	0.91 ± 0.06	5.62 ± 3.24	$0.83 \pm 0.07^{**}$	$6.43 \pm 5.11^{**}$	$0.81 \pm 0.05^{**}$	$6.19 \pm 1.14^{**}$
	M3	$\mathbf{0.92 \pm 0.05}$	$\mathbf{4.67 \pm 3.01}$	$\mathbf{0.87 \pm 0.02}$	$\mathbf{3.91 \pm 0.98}$	$\mathbf{0.85 \pm 0.04}$	$\mathbf{5.00 \pm 1.32}$
	M4	$0.87 \pm 0.08^{**}$	$7.84 \pm 4.13^{**}$	$0.85 \pm 0.04^{**}$	$6.07 \pm 4.65^{*}$	$0.80 \pm 0.08^{**}$	$7.09 \pm 3.26^{**}$
	M5	$0.87 \pm 0.07^{**}$	$9.10 \pm 4.68^{**}$	$0.86 \pm 0.04^{*}$	4.96 ± 3.33	$0.80 \pm 0.10^{*}$	$7.33 \pm 4.24^{*}$
CEM	M1	$\mathbf{0.91 \pm 0.04}$	$6.04 \pm 2.88^{*}$	0.83 ± 0.06	$6.53 \pm 3.69^{**}$	0.82 ± 0.08	6.91 ± 3.08
	M3	$\mathbf{0.91 \pm 0.04}$	5.29 ± 2.63	$\mathbf{0.84 \pm 0.05}$	$\mathbf{5.34 \pm 2.27}$	0.83 ± 0.07	6.35 ± 2.64
	M4	$\mathbf{0.91 \pm 0.04}$	$\mathbf{5.16 \pm 1.99}$	0.83 ± 0.05	5.50 ± 2.24	$\mathbf{0.85 \pm 0.05}^{**}$	$\mathbf{5.51 \pm 1.87}^{**}$
	M5	$\mathbf{0.91 \pm 0.04}$	5.22 ± 2.26	$\mathbf{0.84 \pm 0.05}$	5.34 ± 2.28	$0.84 \pm 0.06^{*}$	$6.03 \pm 2.26^{*}$

F. DIR and nnU-Net comparison

This section discusses the qualitative evaluation between the DIR-based contours (see section II B) and the nnU-Net-based contours obtained with the M3 model (see section II F). Figure 8 plots the histograms of the quality scores assigned by the expert for the contours for each structure, and table V summarizes the mean and standard deviation of these scores (see grading scheme table III). The stars next to the values indicate the statistical significance of the difference between the scores for the DIR-based and nnU-Net-based contours for each structure, based on the p-value computed with the Wilcoxon signed-rank test.

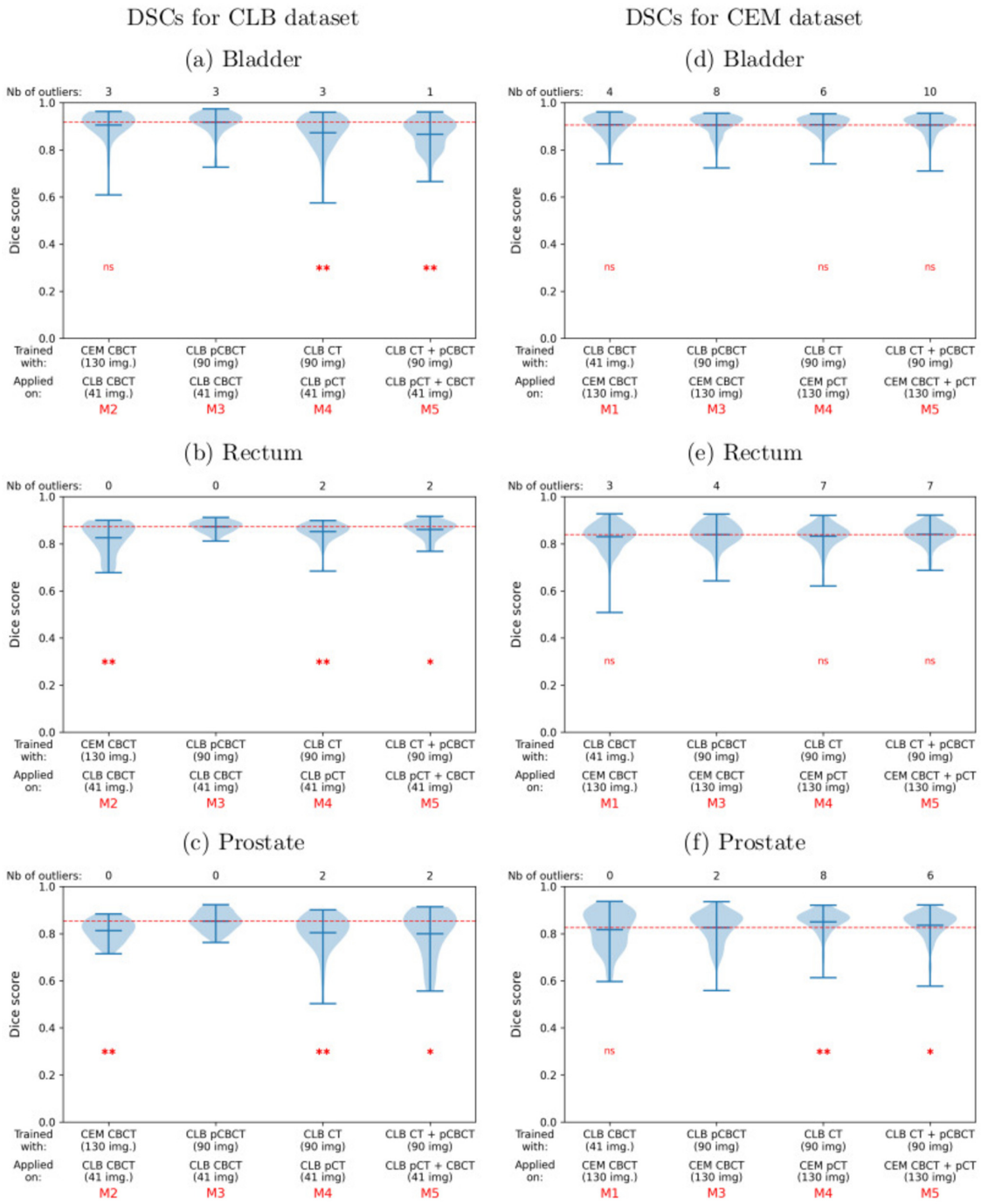

FIG. 6: 6 violin plot panels of the DSCs for the bladder, rectum, and prostate contours for the different models evaluated with the CLB dataset (left) and CEM dataset (right). The dashed red line represents the mean DSC for M3. The statistical significance of the difference between the results of that model and those of M3 is displayed below each violin plot. The "ns" represents no statistical difference, one star represents a statistically significant difference, and two stars represent a highly statistically significant difference.

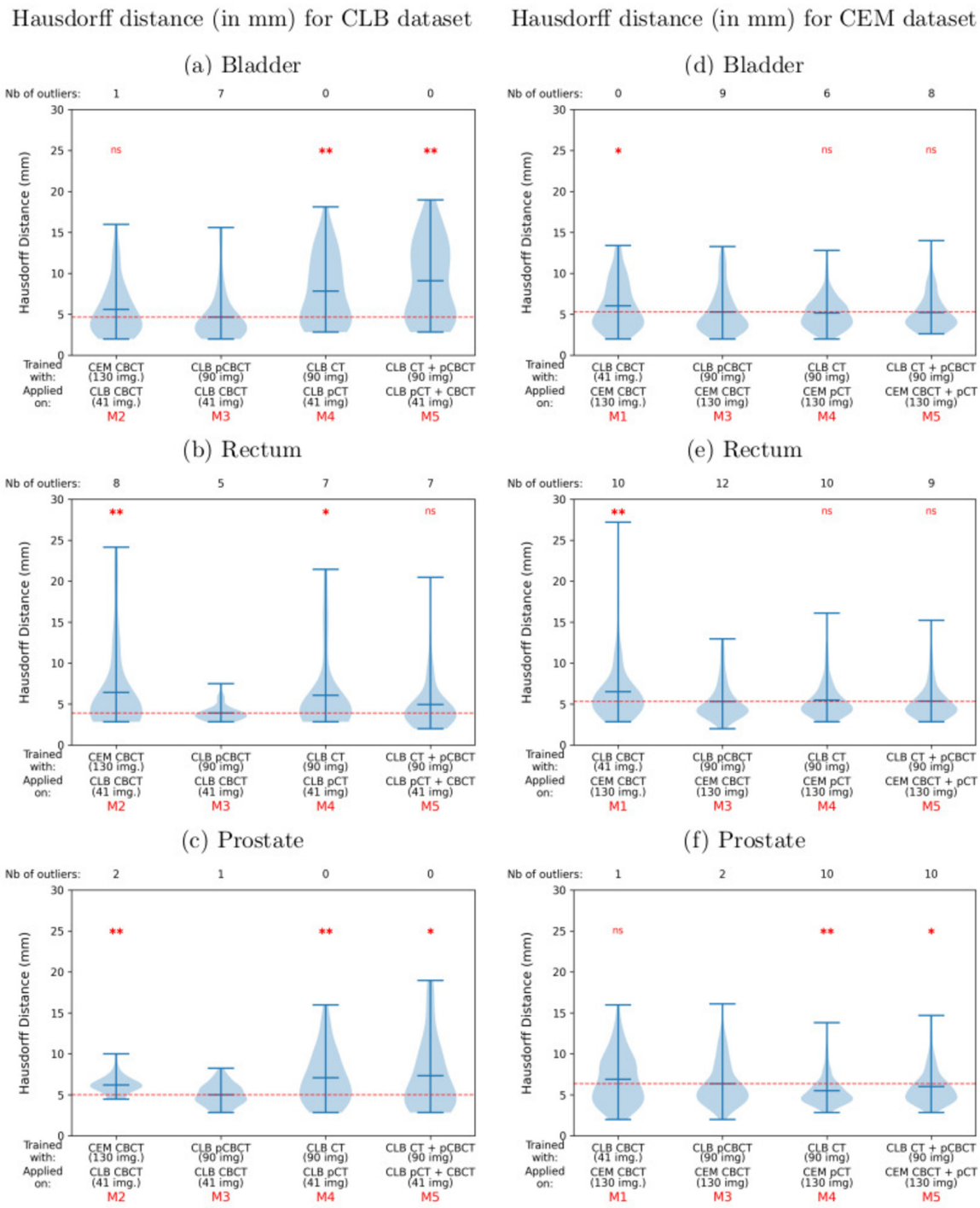

FIG. 7: 6 violin plot panels of Hausdorff distance for the bladder, rectum, and prostate contours for the different models for the CLB dataset (left) and CEM dataset (right). The dashed red line represents the mean Hausdorff distance for M3. The statistical significance of the difference between the results of that model and those of M3 is displayed above each violin plot. The "ns" represents no statistical difference, one star represents a statistically significant difference, and two stars represent a highly statistically significant difference.

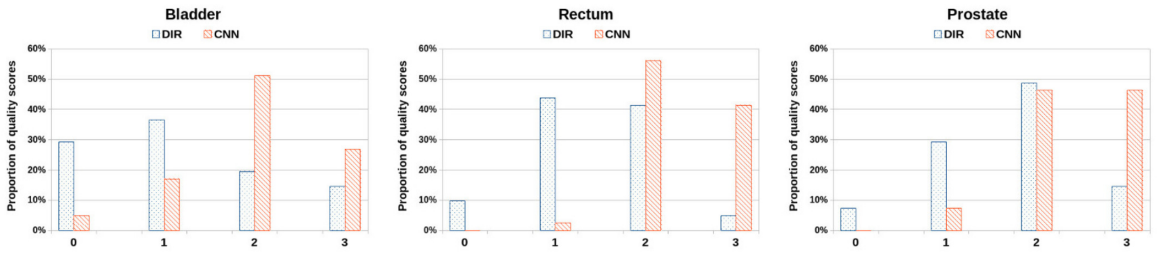

FIG. 8: The distribution of the quality scores of the qualitative evaluation for the DIR-based and nnU-Net-based segmentations.

TABLE V: The mean and standard deviation of the quality scores of the qualitative evaluation for the DIR-based and nnU-Net-based segmentations. The best score per structure is marked with bold letters. A star next to the value in bold indicates a statistically significant difference between the DIR and nnU-Net scores for that structure. Two stars indicate a highly statistically significant difference.

	Bladder	**Rectum**	**Prostate**
DIR	1.20 ± 1.03	1.41 ± 0.74	1.71 ± 0.81
nnU-Net	$\mathbf{2.00 \pm 0.81}^{*}$	$\mathbf{2.39 \pm 0.54}^{**}$	$\mathbf{2.39 \pm 0.63}^{**}$

Figure 9 shows the DIR-based labels (in lighter shades) and nnU-Net-based labels (in darker shades) for three different CLB patients. The quality score for each organ segmentation is also displayed.

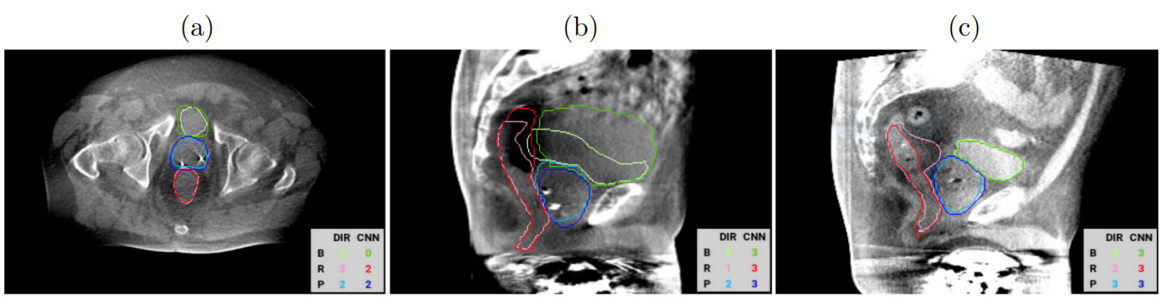

FIG. 9: DIR-based (light) and nnU-Net-based (dark) segmentations for bladder (green), rectum (red), and prostate (blue). The quality scores for the segmentations are also displayed.

Table VI summarizes: 1. the percentage of scores equal or higher than 2 for each structure (criteria 1 in the clinical acceptability evaluation described in section II G), and 2. the percentage of patients whose structures all scored 2 or higher (criteria 2 in the clinical acceptability evaluation).

TABLE VI: The percentage of scores equal or higher than 2 for the DIR- and nnU-Net-based segmentations per structure and combined. For the combined calculation, all structures per patient need to score 2 or higher.

	Bladder	**Rectum**	**Prostate**	**Combined**
DIR	34%	46%	63%	29%
nnU-Net	78%	98%	93%	76%

G. Computation times

The CBCT simulations were performed on a cluster of CPUs (IN2P3 Computing Center, CNRS, Lyon, France) and the deep learning tasks on a cluster of GPU (Jean Zay, CNRS, Orsay, France) with NVIDIA V100 SXM2 32 GB. The approximate computation times are summarized in Table VII. The duration of the deterministic (primary) part of the pCBCT simulation takes between 5 to 8 hours on the IN2P3 cluster per image. For the Monte Carlo (secondary) part of the simulations, the number of particles used was set (300,000 particles over 300 jobs) in order to finish at the same time as the primary part. The image reconstruction takes 2 minutes per image. The training of the cycleGAN for the pCT generation takes 24 hours on a GTX-1080-Ti GPU, while the inference takes 10 seconds per image. For the nnU-Net model, the training takes up to 9 hours on the Jean Zay GPU using 40 cores, and the inference takes approximately 10 seconds per image.

TABLE VII: Approximate computation times.

Task	Job	Time	Tool
pCBCT simulation	Simulation job (primary)	5 - 8 hr	Gate on IN2P3 cluster (1 core)
	Simulation job (scatter)	5 - 8 hr	Gate on IN2P3 cluster (1 core)
	Reconstruction	2 min	RTK on IN2P3 cluster (1 core)
pCT generation	CycleGAN training	24 hr	GTX-1080-Ti GPU
	Inference	10 sec	GTX-1080-Ti GPU
CBCT segmentation	nnU-Net training	9 hr	nnU-Net on JeanZay GPU (40 cores)
	Inference	10 sec	nnU-Net on JeanZay GPU (10 cores)

IV. DISCUSSION

Globally, we observed qualitatively (figure 5) and quantitatively (figure 6 & 7) that training on pCBCT performed better or was not statistically different than the other methods, and, in general, with less outliers. Combining training on both pCBCT and CT led to a degradation of the results.

A. pCBCT quality

While GATE simulation and RTK reconstruction aim to replicate realistic CBCT acquisition conditions, there are some residual differences in the pixel intensities between real and simulated CBCT images, as seen in figure 2(d,e). Indeed, during the reconstruction of a real CBCT image, the Elekta system applies additional steps to reduce the effect of scatter, which have not been implemented in our reconstruction algorithm. Moreover, the real CBCT images were acquired using a bowtie filter, which is not implemented in the GATE simulations. Finally, motion during the acquisition of the real CBCT projections induces artefacts which are not simulated. The anatomy of the patient can sometimes differ between the acquisition of the CT and of the CBCT. Nevertheless, the overall image characteristics seem well reproduced and the normalization process in nnU-Net should compensate the differences in pixel range.

B. CLB and CEM comparison

As seen in table II, the test dataset contains images from two hospitals with two slightly different ways to contour organs, e.g. rectums were contoured on more slices in $DB_{CBCT,\,CEM}$ compared to $DB_{CBCT,\,CLB}$. Therefore, when the training is performed with CT from DB_{CT} , one expects slightly better results with the validation dataset from the same hospital (CLB) than with the other hospital (CEM). Indeed, in figures 6 and 7, for a given method (M3, M4 or M5), DSCs and HD were slightly better when validated with cohort CLB (left) than with cohort CEM (right).

C. pCBCT and real CBCT comparison

Figures 6 and 7 also show that the model trained with simulated pCBCT images (M3) generally display a better performance than models trained with real CBCT images (M1 and M2, first column in all panels), or a statistically non-significant difference like the case of the bladder for CLB dataset, which indicates that we are not losing accuracy when using simulated instead of real images.

In general, the performance of one model with respect to the M3 model and the confidence level in this difference is consistent between the DSCs and HD, except for the bladder and rectum contours for the CEM dataset where M1 has a statistically significant lower performance than M3 in figure 7, but the M1 DSCs are statistically not different from the M3 DSCs. But even those two cases do not contradict the conclusion that M3 generally performs as good, and often better than M1 and M2.

D. pCBCT and pCT comparison

1. CLB dataset

The violin plot panels also show that the model trained with simulated pCBCT images (M3) performs better, for the CLB dataset, than the model trained with real CT images and evaluated on pseudo-CT (M4), and that this difference is highly statistically significant for the DSCs and Hausdorff distance for all contours. For example, M3 / M4 scored mean DSCs of 0.92 ± 0.05 / 0.87 ± 0.08 for the bladder, 0.87 ± 0.02 / 0.85 ± 0.04 for the rectum,

and 0.85 ± 0.04 / 0.80 ± 0.08 for the prostate. M3 also has a lower number of outliers than M4 for the DSCs, but it has a higher number of outliers for the HD for the bladder and the prostate. Nonetheless, this does not indicate a worse performance in these cases, because it can be seen from the distribution of the values for M3 and M4 in the violin plots of the left panels of figure 7 that the HD values for M3 are more concentrated towards the lower values (thicker violin in the bottom), which indicates a lower IQR in these cases which increases the sensitivity to outliers.

2. CEM dataset

For the CEM dataset, M3 performed better than M4 for the rectum contour, however, this difference is not statistically significant. For the bladder and prostate contours, M3 has a lower performance than M4, but the difference for the bladder contour is not statistically significant (M4 column in the top panel on the right in figures 6 and 7), while the difference for the prostate contour is highly significant (M4 column in the bottom panel on the right in figures 6 and 7). It may be related to the difference in the way the prostate is contoured at CEM compared to CLB, and how the high inter-observer variability for the prostate in CT and CBCT images affects the certitude of evaluation metrics such as DSC and HD comparisons[19]. Rectum and bladder DSC and HD values fluctuate less because of their large sizes, since the main source of errors is at the boundary of the organs[19].

E. pCBCT and multi-modal images comparison

We observe that including both types of images as channels (M5) does not improve the performance in comparison to the models that use one modality (M3 and M4). Indeed, M5 tends to perform like M3 and M4 with lower DSCs and higher Hausdorff distance. For example for the bladder contour for the CLB dataset, M3 scored mean DSC of 0.92 ± 0.05, M4 scored 0.87 ± 0.08, and M5 scored 0.87 ± 0.07. For the CLB dataset, it can be seen in the left panels of figures 6 and 7 that M5 always has a lower performance than M3 for all contours, and these differences are all statistically significant. For the CEM dataset (panels on the right in figures 6 and 7), M5 performs as good, and sometimes even better than M3 for the DSCs and HD for all contours. The shape of the violin plots and the number of

outliers are close between these models for the bladder and rectum contours. However, for the prostate contour, M5 has a higher number of outliers in both figures 6 and 7, but a better-shaped distribution where the values are more concentrated towards the higher values in the DSC, and the lower values in the Hausdorff distance. These differences between the results of those two models are statistically non-significant, except for the results of the prostate contour. So M4 and M5 only improved the scores of M3 for the prostate contour for the CEM dataset, while they showed a lower performance or a statistically non-significant difference in all other cases. The results for the CEM dataset indicate that the method still require improvement in order to accurately segment images across hospitals. Indeed, we hypothesize that the differences in contouring (inter-expert variability) are larger than differences in the auto-contour methods.

F. nnU-Net and DIR comparison

Regarding M3 performance in comparison to a DIR-based method, we see in figure 8 that nnU-Net-based contours tend to score 2 or higher (only minor or no corrections needed) more often than DIR-based contours. This can also be seen in the mean of these scores in table V, where DIR-based / nnU-Net-based contours scored an average of 1.20 ± 1.03 / 2.00 ± 0.81, 1.41 ± 0.74 / 2.39 ± 0.54, 1.71 ± 0.81 / 2.39 ± 0.63 for the bladder, rectum, and prostate contours respectively. These differences are also found to be statistically significant (bladder) or highly significant (rectum and prostate) by the Wilcoxon signed-rank test. Figure 9-b also shows the limitations of DIR methods in accounting for bladder deformation. Table VI shows that for DIR-based segmentations, 34% of bladder segmentations, 46% of rectum segmentations, and 63% of prostate segmentations required minor or no corrections, while those percentages are considerably higher for the nnU-Net-based scores, where 78% of bladder segmentations, 98% of rectum segmentations, and 93% of prostate segmentations required minor or no corrections. Similarly, for 29% of the patients, all structures of the patient received a score of 2 or higher for the DIR-based segmentations, while that percentage rises to 76% for nnU-Net-based segmentations. So for M3, both criteria for clinical acceptability are fulfilled, except for the bladder contour where less than 80% (78%) of the patients received a score of 2 or higher. Nonetheless, the mean score for that structure was 2.00 ± 0.81. And so overall, these results are encouraging for the adoption of automated nnU-Net-based segmentation into

the clinical workflow. This qualitative evaluation provides a clinically-oriented comparison between the segmentation methods, and is better able to take into account the differences at the boundaries of the organs than DSC or HD calculations. However, while the process was blinded, it may be sensitive to the scorer's cognitive bias, since only one expert evaluated the segmentations and marked the scores.

G. Related studies

In another study[17,36], the authors presented a data augmentation method that generates multiple CBCTs from a single deformably registered baseline CBCT and planning CT pair, by extracting artifacts from the CBCT and adding them to the corresponding pCT. The resulting synthetic CBCTs are then used in the training of a deep learning model for CBCT segmentation. This can be related to the method proposed in this paper as a training dataset of paired CT/CBCT images is created, but, in our case, we opted for generating CBCT using Monte Carlo method as it is known to be the most accurate way to reproduce the physical effects of CBCT image acquisition. Moreover, our approach does not require initial deformable registered planning CT and week1 CBCT pair, because any planning CT image can be converted into its CBCT counterpart (hence perfectly registered). However, Monte Carlo is a slow method and other methods could be investigated to create realistic CBCT, e.g. Generative Adversarial Net (GAN). In their latest work[17], Dahiya et al. used an image-to-image translation method based on conditional generative adversarial networks (cGANs) for segmenting and translating CBCT to CT at the same time. Their results compare to the state-of-the-art results for these tasks, however, one challenge is to develop a loss function that reduces GAN-produced artifacts without sacrificing the segmentation results for the smaller structures. In our case, we choose to use a simple and robust U-Net method, focusing on the impact of the type of images used in the training dataset.

V. CONCLUSION

The aim of this work was to investigate the interest of using pseudo-CBCT images simulated from CT images for the training of U-Net deep learning model for pelvic CBCT segmentation. This approach avoids the use of contours on real CBCT images that are dif-

ficult to obtain and is an alternative to the use of pseudo-CT images computed from CBCT images and to DIR-based segmentations. Comparison of the Dice and Hausdorff scores shows that the nnU-Net trained with pCBCT performs equally or better on almost all the evaluated test sets. Qualitative evaluation shows the clinical advantage of nnU-Net-based segmentations over DIR-based ones. Contouring variability between different hospitals plays a role as results were slightly lower when models were trained from contour data from one hospital and tested with contours performed by different physicians in another hospital.

Acknowledgements

This work was performed as part of the DELPEL project with financial support from ITMO Cancer AVIESAN (Alliance Nationale pour les Sciences de la Vie et de la Santé/ National Alliance for Life Sciences & Health) within the framework of the Cancer Plan. This work was granted access to the HPC resources of IDRIS under the allocation made by GENCI (Jean Zay computing center). This work was partly performed within the framework of the SIRIC LYriCAN Grant INCa-INSERM-DGOS-12563, and the LABEX PRIMES (ANR-11-LABX-0063) of Université de Lyon, within the program "Investissements d'Avenir" (ANR-11-IDEX-0007) operated by the ANR.

Conflict of Interest Statement

The authors have no relevant conflicts of interest to disclose.

References

[1] M. Nassef, A. Simon, G. Cazoulat, A. Duménil, C. Blay, C. Lafond, O. Acosta, J. Balosso, P. Haigron, and R. de Crevoisier, "Quantification of dose uncertainties in cumulated dose estimation compared to planned dose in prostate IMRT," Radiotherapy and Oncology **119**, 129–136 (Apr. 2016).

[2] S. J. Gardner, N. W. Wen, J. Kim, C. Liu, D. Pradhan, I. Aref, R. Cattaneo, S. M. Vance, B. Movsas, I. J. Chetty, and M. Elshaikh, "Contouring variability of human- and deformable-generated contours in radiotherapy for prostate cancer.," Physics in medicine and biology **60** **11**, 4429–47 (2015).

[3] M. Thor, J. B. B. Petersen, L. Bentzen, M. Høyer, and L. P. Muren, "Deformable image registration for contour propagation from CT to cone-beam CT scans in radiotherapy of prostate cancer," Acta Oncologica **50**, 918 – 925 (2011).

[4] O. Acosta, J. Dowling, G. Drean, A. Simon, R. d. Crevoisier, and P. Haigron, "Multi-atlas-based segmentation of pelvic structures from CT scans for planning in prostate cancer radiotherapy," Abdomen and Thoracic Imaging 623–656, Springer US (Nov 2013).

[5] V. Zambrano, H. Furtado, D. Fabri, C. Lütgendorf-Caucig, J. Gora, M. Stock, R. Mayer, W. Birkfellner, and D. Georg, "Performance validation of deformable image registration in the pelvic region," Journal of Radiation Research **54**, i120–i128 (07 2013).

[6] A. J. Woerner, M. Choi, M. M. Harkenrider, J. C. Roeske, and M. Surucu, "Evaluation of deformable image registration-based contour propagation from planning CT to cone-beam CT," Technology in Cancer Research & Treatment **16**, 801–810 (2017). PMID: 28699418.

[7] P. Meyer, V. Noblet, C. Mazzara, and A. Lallement, "Survey on deep learning for radiotherapy," Computers in Biology and Medicine **98**, 126–146 (2018).

[8] S. Kazemifar, A. Balagopal, D. Nguyen, S. McGuire, R. Hannan, S. Jiang, and A. Owrangi, "Segmentation of the prostate and organs at risk in male pelvic CT images using deep learning," (2018).

[9] X. Liang, H. Morgan, D. Nguyen, and S. Jiang, "Deep learning based CT-to-CBCT deformable image registration for autosegmentation in head and neck adaptive radiation therapy," (2021).

[10] M. Sun and J. Star-Lack, "Improved scatter correction using adaptive scatter kernel superposition," Physics in Medicine & Biology **55**, 6695 (2010).

[11] T. Niu, M. Sun, J. Star-Lack, H. Gao, Q. Fan, and L. Zhu, "Shading correction for on-board cone-beam CT in radiation therapy using planning MDCT images," Medical physics **37**, 5395–5406 (2010).

[12] U. Stankovic, L. S. Ploeger, M. van Herk, and J.-J. Sonke, "Optimal combination of anti-scatter grids and software correction for CBCT imaging," Medical physics **44**, 4437–4451 (2017).

[13] S. Kida, S. Kaji, K. Nawa, T. Imae, T. Nakamoto, S. Ozaki, T. Ohta, Y. Nozawa, and K. Naka-

gawa, "Visual enhancement of cone-beam CT by use of cyclegan," Medical physics **47**, 998–1010 (2020).

[14] H. Sun, R. Fan, C. Li, Z. Lu, K. Xie, X. Ni, and J. Yang, "Imaging study of pseudo-CT synthesized from cone-beam CT based on 3d cyclegan in radiotherapy.," Frontiers in Oncology **11**, 603844–603844 (2021).

[15] J. Zhao, Z. Chen, J. Wang, F. Xia, J. Peng, Y. Hu, W. Hu, and Z. Zhang, "Mv CBCT-based synthetic CT generation using a deep learning method for rectal cancer adaptive radiotherapy," Frontiers in Oncology **11**, 1733 (2021).

[16] X. Dai, Y. Lei, J. Wynne, J. Janopaul-Naylor, T. Wang, J. Roper, W. J. Curran, T. Liu, P. Patel, and X. Yang, "Synthetic CT-aided multiorgan segmentation for CBCT-guided adaptive pancreatic radiotherapy," Medical Physics **48**, 7063–7073 (2021).

[17] N. Dahiya, S. R. Alam, P. Zhang, S.-Y. Zhang, T. Li, A. Yezzi, and S. Nadeem, "Multitask 3D CBCT-to-CT translation and organs-at-risk segmentation using physics-based data augmentation," Medical Physics **48**, 5130–5141 (2021).

[18] J. Jiang, S. Riyahi Alam, I. Chen, P. Zhang, A. Rimner, J. O. Deasy, and H. Veeraraghavan, "Deep cross-modality (MR-CT) educed distillation learning for cone beam CT lung tumor segmentation," Medical Physics **48**, 3702–3713 (2021).

[19] J. Schreier, A. Genghi, H. Laaksonen, T. Morgas, and B. Haas, "Clinical evaluation of a full-image deep segmentation algorithm for the male pelvis on cone-beam CT and CT," Radiotherapy and Oncology **145**, 1–6 (2020).

[20] C. Boydev, D. Pasquier, F. Derraz, L. Peyrodie, A. Taleb-Ahmed, and J.-P. Thiran, "Automatic prostate segmentation in cone-beam computed tomography images using rigid registration," 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) 3993–3997 (2013).

[21] B. Rigaud, A. Simon, J. Castelli, C. Lafond, O. Acosta, P. Haigron, G. Cazoulat, and R. de Crevoisier, "Deformable image registration for radiation therapy: principle, methods, applications and evaluation," Acta Oncologica **58**, 1225–1237 (2019). PMID: 31155990.

[22] D. Sarrut, M. Bardiès, N. Boussion, N. Freud, S. Jan, J.-M. Létang, G. Loudos, L. Maigne, S. Marcatili, T. Mauxion, P. Papadimitroulas, Y. Perrot, U. Pietrzyk, C. Robert, D. R. Schaart, D. Visvikis, and I. Buvat, "A review of the use and potential of the GATE monte carlo simulation code for radiation therapy and dosimetry applications," Medical physics **41**, 064301 (Jun. 2014).

[23] D. Sarrut, M. Bała, M. Bardiès, J. Bert, M. Chauvin, K. Chatzipapas, M. Dupont, A. Etxebeste, L. M. Fanchon, S. Jan, G. Kayal, A. S. Kirov, P. Kowalski, W. Krzemien, J. Labour, M. Lenz, G. Loudos, B. Mehadji, L. Ménard, C. Morel, P. Papadimitroulas, M. Rafecas, J. Salvadori, D. Seiter, M. Stockhoff, E. Testa, C. Trigila, U. Pietrzyk, S. Vandenberghe, M.-A. Verdier, D. Visvikis, K. Ziemons, M. Zvolský, and E. Roncali, "Advanced monte carlo simulations of emission tomography imaging systems with GATE," Physics in Medicine & Biology **66**, 10TR03 (may 2021).

[24] G. Vilches-Freixas, J. Létang, S. Brousmiche, E. Romero, M. Vila Oliva, D. Kellner, H. Deutschmann, P. Keuschnigg, P. Steininger, and S. Rit, "Technical note: Procedure for the calibration and validation of kilo-voltage cone-beam CT models," Med Phys **43**, 5199–5204 (2016).

[25] G. Poludniowski, P. Evans, V. Hansen, and S. Webb, "An efficient monte carlo-based algorithm for scatter correction in keV cone-beam CT," Physics in medicine and biology **54**, 3847–64 (Jul. 2009).

[26] C. Zöllner, "Investigation of a projection scatter correction algorithm for x-ray cone beam computed tomography," Master thesis, Ludwig Maximilians Universität München (Dec. 2016).

[27] S. Rit, M. V. Oliva, S. Brousmiche, R. Labarbe, D. Sarrut, and G. C. Sharp, "The reconstruction toolkit (RTK), an open-source cone-beam CT reconstruction toolkit based on the insight toolkit (ITK)," Journal of Physics: Conference Series **489**, 012079 (Mar. 2014).

[28] Y.-K. Park, G. C. Sharp, J. Phillips, and B. A. Winey, "Proton dose calculation on scatter-corrected CBCT image: Feasibility study for adaptive proton therapy," Medical Physics **42**, 4449–4459 (Aug. 2015).

[29] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2242–2251 (2017).

[30] M. Eckl, L. Hoppen, G. Sarria, J. Boda-Heggemann, A. Simeonova-Chergou, V. Steil, F. Giordano, and J. Fleckenstein, "Evaluation of a cycle-generative adversarial network-based cone-beam CT to synthetic CT conversion algorithm for adaptive radiation therapy," Physica Medica **80**, 308–316 (11 2020).

[31] M. F. Spadea, M. Maspero, P. Zaffino, and J. Seco, "Deep learning based synthetic-CT generation in radiotherapy and PET: A review," Medical Physics **48**, 6537–6566 (2021).

[32] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," Nature methods **18**, 203—211 (February 2021).

[33] F. Isensee, J. Petersen, A. Klein, D. Zimmerer, P. F. Jaeger, S. Kohl, J. Wasserthal, G. Koehler, T. Norajitra, S. Wirkert, *et al.*, "nnU-Net: Self-adapting framework for U-Net-based medical image segmentation," arXiv preprint arXiv:1809.10486 (2018).

[34] "MedPy's documentation." http://loli.github.io/medpy/, (Feb. 14 2019). Accessed: 2021-02-04.

[35] "SciPy's documentation." https://docs.scipy.org/doc/scipy/, (Aug. 10 2021). Accessed: 2021-12-14.

[36] S. R. Alam, T. Li, P. Zhang, S.-Y. Zhang, and S. Nadeem, "Generalizable cone beam CT esophagus segmentation using physics-based data augmentation," Physics in Medicine & Biology **66**, 065008 (mar 2021).

List of Figures

List of Tables