Computers in Biology and Medicine 131 (2021) 104269

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: http://www.elsevier.com/locate/compbiomed

Domain adversarial networks and intensity-based data augmentation for male pelvic organ segmentation in cone beam CT

Eliott Brion [a,*], Jean Léger [a], A.M. Barragán-Montero [c], Nicolas Meert [b], John A. Lee [a,c], Benoit Macq [a,**]

[a] ICTEAM, UCLouvain, Louvain-la-Neuve, 1348, Belgium
[b] Hôpital André Vésale, Montigny-le-Tilleul, 6110, Belgium
[c] IREC/MIRO, UCLouvain, Brussels, 1200, Belgium

ARTICLE INFO

Keywords:
Deep learning
Radiotherapy
Unsupervised domain adaptation
Segmentation

ABSTRACT

In radiation therapy, a CT image is used to manually delineate the organs and plan the treatment. During the treatment, a cone beam CT (CBCT) is often acquired to monitor the anatomical modifications. For this purpose, automatic organ segmentation on CBCT is a crucial step. However, manual segmentations on CBCT are scarce, and models trained with CT data do not generalize well to CBCT images. We investigate adversarial networks and intensity-based data augmentation, two strategies leveraging large databases of annotated CTs to train neural networks for segmentation on CBCT. Adversarial networks consist of a 3D U-Net segmenter and a domain classifier. The proposed framework is aimed at encouraging the learning of filters producing more accurate segmentations on CBCT. Intensity-based data augmentation consists in modifying the training CT images to reduce the gap between CT and CBCT distributions. The proposed adversarial networks reach DSCs of 0.787, 0.447, and 0.660 for the bladder, rectum, and prostate respectively, which is an improvement over the DSCs of 0.749, 0.179, and 0.629 for "source only" training. Our brightness-based data augmentation reaches DSCs of 0.837, 0.701, and 0.734, which outperforms the morphons registration algorithms for the bladder (0.813) and rectum (0.653), while performing similarly on the prostate (0.731). The proposed adversarial training framework can be used for any segmentation application where training and test distributions differ. Our intensity-based data augmentation can be used for CBCT segmentation to help achieve the prescribed dose on target and lower the dose delivered to healthy organs.

1. Introduction

About half of cancer patients are currently treated with radiotherapy. This treatment consists in delivering a prescribed dose of radiation, typically X-rays, to the tumor, while limiting the dose delivered to the surrounding healthy organs, also called organs at risk or OARs. The spatial distribution of the dose is optimized at treatment planning, which requires a computed tomography (CT) scan of the patient and the delineation of the tumor and OAR contours on the scan in order to evaluate the dose delivered to those regions. In current practice, trained clinicians delineate the contours manually on the CT. Then, the computed dose is split into different fractions, which are delivered to the patient during successive treatment sessions. However, variations in patient anatomy between sessions can impair dose conformity, with typically too large a dose delivered to the healthy organs or too low a dose to the tumor [47]. Those daily variation are particularly large in the pelvic region due to physiological function (e.g., bladder and rectal filling or voiding). This can jeopardize treatment optimality, with possibly lower tumor control probability or more acute secondary effects, affecting quality of life, such as anal canal bleeding for prostate cancer.

Adaptive radiation therapy is aimed at ensuring an optimal treatment by using daily images to monitor anatomical modifications and adapt the treatment plan accordingly. For this purpose, drawing new contours on this daily image is a crucial step [20,49]. Since drawing those contours manually takes too long to be carried out before each treatment session (2–4 h per volumetric image, with one image for each of the ~20 treatment sessions), an automated segmentation algorithm is

* Corresponding author. ELEN, Place du Levant 2/L5.04.04, 1348, Louvain-la-Neuve, Belgium.
** Corresponding author.
E-mail address: eliott.brion@uclouvain.be (E. Brion).

https://doi.org/10.1016/j.compbiomed.2021.104269
Received 7 October 2020; Received in revised form 7 February 2021; Accepted 8 February 2021
Available online 16 February 2021

needed. Registration-based segmentation approaches are widely used for medical image segmentation due to their easy interpretation and high performance in regions with little anatomical variations [46,50]. However, these methods fail in the case of large deformations between scans to be registered, as it is often the case in the pelvic region [57,65]. Furthermore, they are rather slow, reaching 1.92 min per image for the ANACONDA algorithm [61] and 8.33 min per image for the diffeomorphic morphons algorithm, which is suited for registering planning and daily images [25].

While deep neural networks such as U-Net provide state-of-the-art performance for medical image segmentation, their use in our application is not direct. Indeed, it consists of a supervised learning model whose internal parameters must be adjusted on a set of already contoured images. Although such databases are available for CT scans, since contouring them has been part of classical clinical routine, the use of helical CT in the treatment room to position the patient is challenging because physical constraints prevent the CT scanner and the dose delivery machine from being in the same place. Therefore, the daily image is often what can be seen as a noisy CT scan: a cone beam CT (CBCT) scan. Since these images are not contoured in clinical routine, there are no (or few) labeled CBCT scans on which a deep learning model can be trained for CBCT segmentation. This raises the question of how to leverage the large database of annotated CT scans in order to build a model able to segment images from a slightly different (see Fig. 1) domain: CBCT scans.

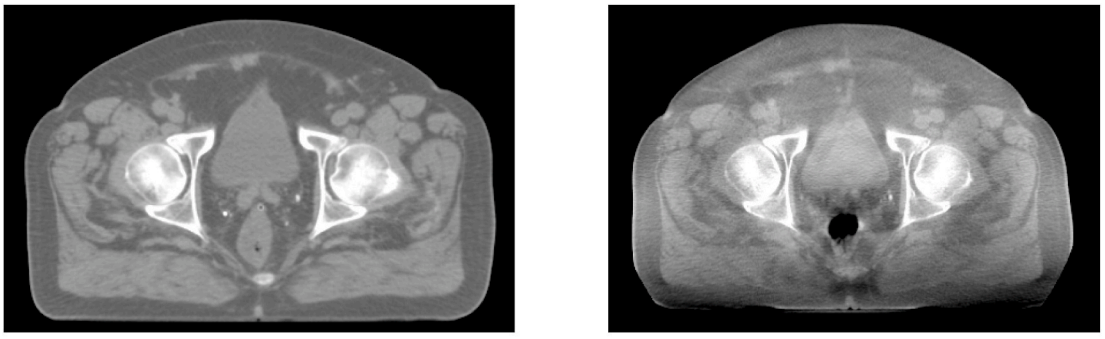
The most direct way to do this is simply to train a network on a mixed training set made of both labeled CT scans and a small number of CBCT scans contoured specifically for model training [5,21,32]. In order to avoid the annotations of CBCT scans, pseudo-CBCT scans can be generated from annotated CT scans [52] and added to the CT scan training set. However, those approaches need either a small number of annotated CBCT scans or a physical model for the pseudo-CBCT scan generation, which requires thorough expertise in the field. In this paper, we report on training a deep neural network for male pelvic organs segmentation in CBCT scans using labeled CT and unlabeled CBCT scans, as available in the current treatment workflow, and limited physical expertise (i.e., it does not require deriving an explicit physical model for CT to CBCT conversion). For this purpose, we propose two different strategies. On the one hand, we follow an unsupervised domain adaptation (UDA) learning strategy where the CT and CBCT scans are considered the source and target domains, respectively. In machine learning, unsupervised domain adaptation refers to strategies that train a model using labeled examples, a “source domain” (in our case, CTs), and only unlabeled examples in a similar yet different “target domain” (in our case, CBCTs). We extend the method proposed in Refs. [18,28] for learning domain-invariant features to a U-Net architecture. This lets us train a segmenter using the annotated CT scans while forcing the features to be transferable to the CBCT scans. On the other hand, we use standard (non adversarial) training with intensity-based data augmentation on the CBCT inputs in order to bring CT and CBCT intensity distributions closer at the input level and enhance the learning of domain-invariant features. Then, the selected best method for intensity-based data augmentation is combined with adversarial training. In addition, the effect of database size on the model’s performance is analyzed for the intensity-based data augmentation approach. Our contribution is twofold:

- Adversarial networks for unsupervised domain adaptation are implemented with 3D U-Net and applied to male pelvic organ segmentation on CBCT, including the bladder, rectum, and prostate. The model is trained using annotated CTs and non-annotated CBCTs, outperforming (for each of the three organs) a model trained only on CT (i.e., without adversarial networks). Our architecture for unsupervised domain adaptation with a gradient reversal layer uses (i) a 3D end-to-end architecture (in contrast to slice-wise 2D ones) and (ii) U-Net as its base (while some previous works use simple autoencoders or patch-based methods, without skip connections). Our code is made public and can be used for any segmentation application experiencing a shift of distribution between training and test images, a recurring problem in biomedical imaging.
- Different strategies for intensity-based data augmentation are compared and we demonstrate that brightness, linear, and bilinear transformations improve generalization of CT-only models to CBCT data without requiring annotated CBCTs. For these strategies, neural networks are trained in a standard, non adversarial fashion. Applying intensity-based data augmentation during training achieves similar results to deformable image registration (DIR) for the prostate and outperforms existing automatic segmentation methods (trained with labeled CTs and unlabeled CBCTs) for the bladder and rectum. Moreover, our model is several orders of magnitude faster than DIR at inference, which is essential for online treatment adaptation. It also outperform a cycleGAN-based approach.

2. Related works

2.1. Deep domain adaptation

While medical imaging datasets are becoming more available, the lack of annotated data (i.e., labels) remains a challenge for supervised machine learning algorithms. Therefore, different approaches have been proposed in order to learn with less supervision. Among those methods, transfer learning considers learning from related learning problems. In particular, domain adaptation is a transfer learning category where the training and test data are from different domains whereas the task



(a) Slice of a CT scan. (b) Slice of a CBCT scan.

Fig. 1. Radiotherapy relies on CT scans for treatment planning and on cone beam CT (CBCT) for treatment monitoring (these can be seen as noisy CTs). Current clinical workflow generates large databases of annotated CT scans on which neural networks can be trained for segmentation tasks. However, using these segmentation networks on CBCT scans at test time requires innovative methods due to differences in appearance between the two modalities.

performed by the model remains similar for both domains [11]. The source domain refers to the domain where labels are available and supervised learning can be performed. The target domain refers to the domain where no or few labels are available and where the actual task should be performed.

When few labels are available in the target domain, supervised learning can be used in both source and target domains. Domain adaptation with fine-tuning strategies has been studied and applied experimentally for the segmentation of brain lesions [19]. Also, cross-domain data augmentation has been proposed in order to segment male pelvic organs in CBCTs. This approach uses labeled CTs to augment the training set [32]. However, such approaches still require some labels in the target domain, which limits their use in applications where target labels are hard to collect. If no labels are available in the target domain, unsupervised domain adaptation (UDA) is considered. We generally divide current UDA methods into three categories: feature-level transferring, image-level transfer, and label-level transfer.

2.2. Feature-level transferring

Feature-level transferring was adopted by the first deep domain adaptation approaches and attempts to force the extraction of domain-invariant features. Those methods extend deep convolutional networks to domain adaptation either (i) by minimizing the maximum mean discrepancy (MMD) between the features learned from the source and target domains with one or more adaptation layers [38,59], (ii) by adding a domain discriminator determining whether the learned deep features are extracted from source or target data following a domain-adversarial learning strategy [1,18,58], or (iii) by minimizing the correlation distance between features [54]. Building domain-invariant features can also rely on disentanglement strategies to build a shared domain-invariant content space and a domain-specific style space [64]. However, feature-level transferring techniques suffer from several limitations, especially when it comes to a segmentation task. First, the semantic consistency is not enforced through the alignment of the marginal distributions [66]. Second, the level of a deep representation to be aligned is not trivial to choose and not easily interpretable, which can lead to a tedious tuning process. Indeed, different choices of feature levels may result in different segmentation results. To tackle those limitations, a significance-aware feature purification before the adversarial adaptation was proposed recently [39], which eases the feature alignment and stabilizes the adversarial training course. Medical image segmentation has been performed using adversarial UDA between MR images [28] and between MR images and CT scans [15]. Adversarial UDA has also been used to segment MR images with both semantic and boundary information [34].

2.3. Image-level transferring

Image-level transferring consists in adapting the source images to look as if they belong to the target domain [3]. Hence, the distribution alignment is performed in the raw pixel space. Domain adaptation can be formulated as the generalization to a new domain. From this model generalization perspective, data augmentation techniques can be used to make the source training samples look closer to the target samples. Image data augmentation for deep learning has been reviewed in Ref. [53] and, with a focus on medical image segmentation in Ref. [43]. Common data augmentation techniques consider affine, elastic and pixel-level transformations. Pixel-level transformation are especially useful for medical image processing since they can simulate the acquisition with different devices. Common transformation are noise injections, modification of the brightness, sharpening or blurring with a kernel and mixup [40,41,44,60]. Such methods are easy to implement and fast but may not be able to produce sufficiently realistic images in order to allow domain adaptation.

In order to cope with those limitations, approaches allowing artificial images generation have been proposed. Image-to-image translation techniques, such as CycleGAN [69], can generate visually appealing results that maintain local content in natural scenes. Image-to-image transferring has been used prior to segment abdominal organs [24], lung tumors [27], cardiac [10,67] and knee [37] structures using domain adaptation between MR images and CT scans. It has also been used for chest X-ray [9] and fundus images [68] segmentation. Besides, a cycle-and shape-consistent GAN has been proposed for segmentation of cardiovascular and abdominal structures in CT scans and MR images, as well as mammography X-rays [6]. However, image-level transferring techniques also have several limitations. Indeed, for image alignment, the quality of the synthetic images is not guaranteed without large amounts of data from the target domain. Furthermore, those approaches tend to correctly deal with pixel or low-level domain shifts but sometimes fail to capture high-level semantic information [23]. Some researchers attempt to overcome the limitations of feature and image alignment strategies by combining them for natural [23] and medical images [10]. In Ref. [10], cardiac substructures are segmented in MR images and CT scans. Alternatively, approaches attempt to improve the shape and appearance of the structures of interest [2,14,27,67]. Recently [26], models the co-dependency between images and their segmentation in order to improve both geometry and appearance consistency through the domain transfer process.

2.4. Label-level transferring

Label-level transferring consists in predicting pseudo-labels on target data and using them to generalize to the target domain. Self-training considers a model trained on the labeled source data to predict labels on the target data, then simply trains on some of those estimated pseudo-labels (e.g., the most confident ones) [30,36,71]. Self-ensembling uses a teacher and a student network, where the teacher model is an ensemble model obtained by averaging the weights of the student model over the training phase. More precisely, the student network is trained such that is produces predictions that are consistent with the teacher predictions on the target data. In this framework, the predictions of the teacher model are considered as pseudo-labels as the teacher model is an ensemble model, supposed to show good generalization capabilities. Even if current self-ensembling techniques are effective in a semi-supervised setting [56], these approaches need much manually-tuned data augmentation in order to succeed in domain alignment tasks [16]. Concerning semantic segmentation, appropriate data augmentation should be performed to avoid a spatial misalignment between the student and teacher predictions and is a limitation of the approach. Medical image segmentation in MR images based on self-ensembling was first performed by Ref. [48], followed by Ref. [45]. In Refs. [12,30], data augmentation is based on a GAN and allows improved semantic segmentation with self-ensembling. Besides self-training and self-ensembling, multi-view co-training has also been used for medical image segmentation and domain adaptation [63].

2.5. CBCT segmentation

Image registration allows contours to be propagated from planning CT to CBCT scans [4,42,55,57,65]. However, those methods fail when large anatomical deformations occur between CT and CBCT scans. Therefore, more sophisticated registration-based approaches have been proposed. This includes deformable image registration (DIR) based on B-splines and mutual information [62]. This last approach uses six successive DIRs with different resolutions and, after visual examination, an optional last pass focusing around the region of interest. Another approach implements a DIR strategy where the deformation is rigidly constrained for bone and/or the prostate, while surrounding tissue can still deform elastically [31].

To eliminate registration-related errors, statistical shape models have been proposed for organ segmentation on CBCT scans [8,51].

However, building such patient-specific shape models require several delineated CBCT scans and landmarks or meshes. More recently, deep learning approaches have been used for the segmentation of CT scans [7,29]. The 3D U-Net fully convolutional neural network has been used with CT and CBCT scans in the training set to segment female [21,22] and male pelvic organs on CBCT scans [5,32]. In order to avoid annotating CBCT scans, pseudo-CBCT scans have been generated from annotated CT scans in order to build a training set including both CT and pseudo-CT scans [52]. In order to improve the segmentation of soft tissues [33], proposed a method to generate synthetic MR images from CBCT scans using a Cycle-GAN network [70]. Then the synthetic MR images are segmented using manual annotations obtained on the original CBCT scans as ground truth. This is extended in Ref. [17], where both CBCT scans and CBCT-generated synthetic MR images have been used to segment male pelvic organs and femoral heads.

3. Materials and methods

In this section, we introduce the dataset and the two main strategies investigated in this study: adversarial networks and intensity-based data augmentation. We also present the metrics used to assess the performance of the different methods, as well as the deep learning and registration baselines.

3.1. Data and preprocessing

In this study, we use data from 134 different patients who underwent radiotherapy for prostate cancer in two different hospitals (41 patients at CHU-Charleroi Hopital André Vésale and 93 patients at CHU-UCLouvain-Namur). For the purposes of the study, we shuffled patients from both hospitals and randomly split them in five cohorts, including four cohorts of 30 patients, denoted $\mathscr{C}_1$, $\mathscr{C}_2$, $\mathscr{C}_3$, $\mathscr{C}_4$ and one cohort of 14 patients, denoted $\mathscr{C}_5$. We collected CT scans for cohorts $\mathscr{C}_1$, $\mathscr{C}_2$, $\mathscr{C}_4$, and $\mathscr{C}_5$ for a total of 104 CT scans. We collected the CBCT scans for cohorts $\mathscr{C}_3$ and $\mathscr{C}_4$ for a total of 60 CBCT scans. The CBCT scans coming from CHU-Charleroi Hopital André Vésale were acquired with a Varian TrueBeam STx version 1.5, while those coming from CHU-UCLouvain-Namur have been acquired with a Varian OBI Cone Beam CT. The use of these retrospective, anonymized data for this study has been approved by both hospitals' ethics committees (May 24, 2017, for CHU-Charleroi Hopital André Vésale and May 12, 2017, for CHU-UCLouvain-Namur). In order to ensure data uniformity across the entire dataset, all the 3D CT and CBCT scans (as well as the 3D binary masks representing the manual segmentations) were re-sampled on a $1.2\times\ 1.2\times\ 1.5$ mm regular grid. All re-sampled image volumes and binary mask volumes were cropped to volumes of $160\times\ 160\times\ 128$ voxels containing the bladder, rectum, and prostate.

The only processing applied to the scans prior to deep learning was normalization by the mean and standard deviation. More precisely, we computed the mean μ_{train} and the standard deviation σ_{train} over all pixels of all CT and CBCT scans used in the training set, indifferently. Before both training and inference, we subtracted from each CT or CBCT scan the mean μ_{train} and then divided by the standard deviation σ_{train}.

3.2. Adversarial networks

The feature-level UDA follows the adversarial learning strategy proposed in Ref. [28] and was implemented using the gradient reversal layer (GRL) approach proposed in Ref. [18]. A GRL is a custom layer for which the gradient is hard-coded to achieve a desired weights update rule (i.e., different from the weight update rule of "standard" back-propagation). A segmenter is trained in the source domain and a domain classifier connected to the intermediate features of the segmenter encourages those features to be domain-invariant. Hence, the segmenter is further split into a feature extractor, which computes the features Lz provided as input to the domain classifier, and a label predictor, which exploits these features, in addition to a few upper skip connections – except for $z=1$ (first layer) and $z=11$ (last layer) – to produce a segmentation mask. The feature extractor is denoted $G_f(\cdot;\theta_f)$, the label predictor, $G_y(\cdot;\theta_y)$, and the domain classifier, $G_d(\cdot;\theta_d)$, where θ_f, θ_y, and θ_d are the weights of the feature extractor, label predictor, and domain classifier, respectively. The segmentation is performed by the composition of the feature extraction and the label prediction. Formally, we define the segmenter as $G_{\text{seg}}(\cdot;\theta_f,\theta_y) = G_y(G_f(\cdot;\theta_f);\theta_y)$. The segmenter is implemented by following a classical U-Net architecture. The proposed architecture is shown in Fig. 2. The features computed at the output of every layer of U-Net are denoted by L1, L2, …, L11. The hyperparameter z defines the layer Lz at which the label predictor and domain classifier are connected to the feature extractor.

Both the segmenter and domain classifier are updated with every batch of data. The first half of the batch contains annotated CT scans from patient cohort $\mathscr{C}_1$ (i.e., source data), whereas the second half of the batch contains unlabeled CBCT scans for patient cohort $\mathscr{C}_3$ (i.e., target data). Note that they are unpaired data (i.e., each batch contains at most one image of any given patient). The annotated CT scans are used to update the weights of the feature extractor and label predictor following a classical end-to-end supersized segmentation setting. Additionally, the CT and CBCT scans are used to train the feature extractor and domain classifier following an adversarial strategy. Following [18,28], this is achieved by finding the saddle point of the following function:

$$L_{\text{tot}}(\theta_f,\theta_y,\theta_d) = \sum_{i=1..Nd^i=0} L_{\text{seg}}\left(\widehat{\mathbf{Y}}^i(\theta_f,\theta_y),\mathbf{Y}^i\right) + \sum_{i=1..2N} L_{\text{dom}}\left(\widehat{d}^i(\theta_f,\theta_d),d^i\right), \tag{1}$$

where $\widehat{d}^i$ is the probability of the ith example to belong to the target domain and N is the number of examples for each domain (i.e., there are $2N$ examples in total). In this loss, L_{seg} is the dice loss and encourages both the feature extractor and the label predictor to learn features useful for source domain image segmentation. The expression L_{dom} is the cross-entropy and encourages (i) the feature extractor to learn domain-invariant features and (ii) the domain classifier to predict whether the features in Lz are activated by a source or target domain input.

The label predictor's output is:

$$\widehat{\mathbf{Y}}^i(\theta_f,\theta_y) = G_y\left(G_f\left(\mathbf{X}^i;\theta_f\right);\theta_y\right), \tag{2}$$

where $\mathbf{X}^i$ is the ith input image. The saddle point of (1) is estimated by backpropagation. To control the strength of the interaction between the feature extractor and the domain classifier, we follow [18] and use the *gradient reversal layer*, defined as $R_\lambda(\mathbf{a}) = \mathbf{a}$ and $\mathrm{d}R_\lambda(\mathbf{a})/\mathrm{d}\mathbf{a} = -\lambda\mathbf{I}$. The larger λ, the higher the interaction between domain classifier and feature extractor. The domain classifier's output then is written as:

$$\widehat{d}^i(\theta_f,\theta_d) = G_d\left(R_\lambda\left(G_f\left(\mathbf{X}^i;\theta_f\right)\right);\theta_y\right). \tag{3}$$

In practice, setting λ to a fixed value throughout learning leads to instabilities. To address this issue, we set λ to zero during the first e_0 epochs (i.e., the feature extractor and domain classifier learn their tasks independently) and then increase λ linearly until it reaches the value $\lambda_{\max}$ after the total number of epochs n_{epochs}:

$$\lambda(e) = \max\left(0,\lambda_{\max}\frac{e-e_0}{n_{\text{epochs}}-e_0}\right), \tag{4}$$

where e is the current epoch. In this expression, $\lambda_{\max}$, e_0 and n_{epochs} are hyper-parameters.

In supervised learning, hyper-parameters are most often chosen by comparing the performance on a validation set for different values of hyper-parameters. When the training involves images from different distributions, the validation set must contain images from the distribution on which we want the system to perform best (i.e., validation set contains annotated images from the target domain). In an unsupervised

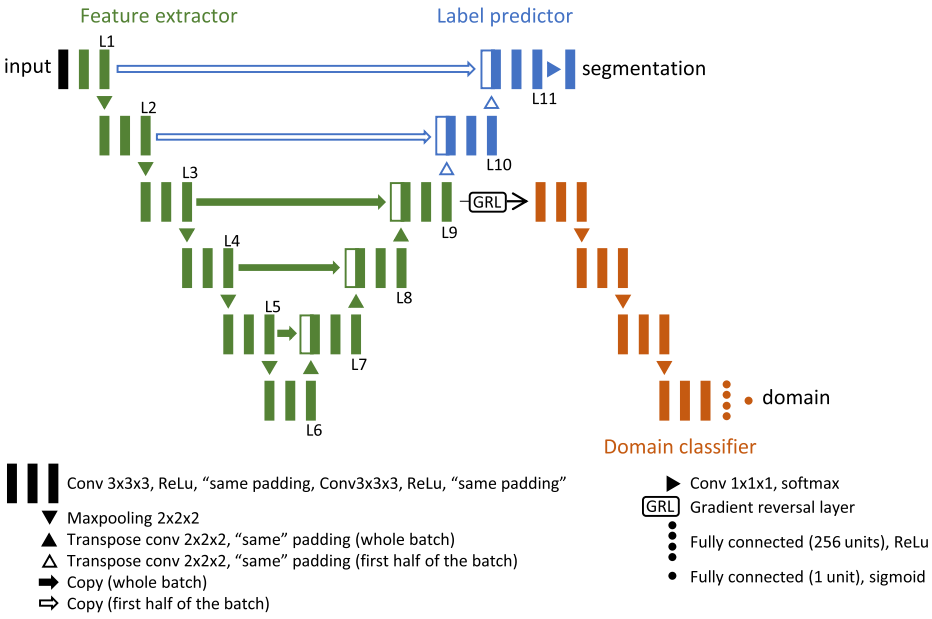


Fig. 2. Our proposed adversarial architecture adapts 3D U-Net to unsupervised domain adaptation by backpropagation (in this figure, the layer Lz, $z = 9$ is adapted). The goal is to learn, from labeled source examples (in our case, CTs), features that are also useful for segmenting images from a different yet similar target domain (CBCTs). U-Net is split in two parts: a feature extractor and a label predictor. The feature extractor learns features from the whole batch (both source and target), with its final layer Lz to (i) be useful for the label predictor and (ii) fool the domain classifier. The label predictor learns feature from half the batch in the layer Lz (the source examples) aiming to predict a segmentation mask for these source examples. The domain classifier uses these same features Lz to classify all images in the batch (both source and target) as being either source or target.

domain adaptation setting, however, we need alternative ways to choose hyper-parameters (since annotations are assumed not to be available for the target domain). In our application, we assess the performance of our method by observing (i) the DSC on the test source data and (ii) the domain classifier accuracy (i.e., the proportion of images correctly classified as CT or CBCT). More specifically, we observe good performance on the target domain when the following conditions[1] are met:

- during the first e_0 epochs, (i.e., when the feature extractor and domain classifier are trained independently), the test domain classifier accuracy steadily reaches a test accuracy close to its theoretical maximum (i.e., its value at convergence when $\lambda = 0$ throughout training);
- during the remaining training, the test domain classifier accuracy decreases to 0.5 and;
- during the whole training, the DSCs of the bladder, rectum, and prostate on the source domain reach values close to their theoretical maxima (its value at convergence when there is no UDA).

We selected hyper-parameter values that best meet these criteria. For the feature extractor and label predictor, our network starts with 16 feature maps in the first level (corresponding to L1) and is doubled in each level. For the domain classifier, the number of feature maps in the first level is the number of feature maps in Lz divided by two, and then it doubles with each increase in level. In the example in Fig. 2, where the domain classifier is connected on L9 (which has 64 feature maps), the domain classifier has 32 feature maps in its first level, followed by 64, 128, and 256 in the following levels. The number of units in the last fully connected layer is always equal to 256. Due to memory limitation, we choose a batch size of 2. Our model is optimized with Adam and a learning rate of 10^{-4}. For the trade-off parameter schedule, a value of $e_0 = 50$ was chosen and $\lambda_{\max}$ was set to 0.01, 0.03, 0.1 or 0.3 depending on the layer adapted. All networks were trained for 150 epochs.

The source code is publicly available on https://github.com/eliottbrion/unsupervised-domain-adaptation-unet-keras.

[1] Note that these conditions do not require annotated target images, so that hyper-parameters can be selected using only annotated source images and non-annotated target images. In this study, the labeled CBCTs (30 images from cohort $\mathscr{C}_4$) are used only to report the method's performance.

3.3. Intensity-based data augmentation

In the context of deep learning, data augmentation most often refers to the application of "geometric" transformations to images (i.e., small rotations, translations, and shear) in the training set in order to improve generalization on the test set. The rationale is that samples from the training and test sets are all generated by the same distribution, and the goal of geometric data augmentation is to mitigate the small number of training samples. However, when considering domain adaptation, training and test samples belong to different distributions. Hence, on top of the geometric data augmentation, we have performed an "intensity-based" data augmentation. The goal is to bring the source distribution closer to the target distribution. The following augmentations have been implemented:

- Brightness: a random offset is added to the image: $\mathbf{X} \leftarrow \mathbf{X} + u\mathbf{J}$, where $u \sim U[-100, 100]$ and $\mathbf{J} \in \mathbb{R}^{160\times160\times128}$ is a tensor with all entries equal to one.
- Contrast: the mean separation between low- and high-intensity voxels of the image is randomly modified: $\mathbf{X} \leftarrow u(\mathbf{X} - \overline{X}\mathbf{J}) + \overline{X}\mathbf{J}$.
- Sharpness: high-frequency detail is randomly strengthened or weakened: $\mathbf{H} = \mathbf{X} - G(\mathbf{X}, 2^2)$ and $\mathbf{X} \leftarrow \mathbf{X} + u(\mathbf{H} - \overline{H}\mathbf{J})$, where $\mathbf{H}$ is a high-frequency image obtained by subtracting a Gaussian blurred (with standard deviation of 2 HU) version of the image and $u \sim U[-0.2, 1.8]$.
- Noise: random Gaussian noise with zero mean and standard deviation 20 HU is added to each voxel of the image: for $1 \le i,j \le 160$ and $1 \le k \le 128$, $X_{i,j,k} \leftarrow X_{i,j,k} + v_{i,j,k}$, where $v_{i,j,k} \sim N(0, 20^2)$.
- Linear: we assume unknown linear correspondence between CT and CBCT intensities: $\mathbf{X} \leftarrow m\mathbf{X} + p\mathbf{J}$, where $m \sim U[0.5, 1.5]$ and $p \sim U[-50, 50]$.
- Bilinear: the correspondence between CT and CBCT intensities are often best modelled by a piecewise linear function instead of a strictly linear function. This hinged line is inspired by typical *calibration curves* from Hounsfield units to other physical quantities. We model this correspondence during training: for $1 \le i,j \le 160$ and $1 \le k \le 128$,

$$X_{i,j,k} \leftarrow \begin{cases} m_1 X_{i,j,k} + p, & \text{if } X_{i,j,k} \le 0 \\ m_2 X_{i,j,k} + p, & \text{otherwise} \end{cases}$$

where $m_1, m_2 \sim U[0.5, 1.5]$ and $p \sim U[-50, 50]$.

To finish, the influence of the data size on the accuracy of the model is also studied by increasing the training size from 30 to 74.

3.4. Performance metrics and comparison baselines

Performance metrics. The performances of the different methods are measured with one overlap-based metric (the dice similarity coefficient, or DSC) and one distance-based metric (the symmetric mean boundary distance, or SMBD). They are respectively defined by:

$$\mathrm{DSC} = \frac{2|M \cap P|}{|M| + |P|}, \tag{5}$$

$$\mathrm{SMBD} = \frac{\overline{D}(M,P) + \overline{D}(P,M)}{2}, \tag{6}$$

where M and P are the sets containing the matricial indices of the manual and predicted segmentation 3D binary masks (respectively), $\overline{D}(M,P)$ is the mean of $D(M,P)$ over the voxels of Ω_M, and $D(M,P) = \{\min_{x \in \Omega_P} ||\mathbf{s} \odot (\mathbf{x} - \mathbf{y})||, \mathbf{y} \in \Omega_M\}$, where Ω_M, Ω_P are respectively the boundaries extracted from M and P and $\mathbf{s}^\top = (1.2, 1.2, 1.5)$ is the pixel spacing in mm.

Deep learning baselines. A naive strategy for learning to segment target domain images from a dataset of labeled source domain images would be simply to train a neural network on those source images (CTs from cohort $\mathscr{C}_1$, $\mathscr{C}_2$, or $\mathscr{C}_5$, depending on the experiment) and then test it on test domain images (always the CBCTs from cohort $\mathscr{C}_4$). The CBCTs from cohort $\mathscr{C}_3$ are not used for tests since the images are already used for training the adversarial networks and we wish to report the performance on an independent test set. This "source only" strategy is a valid baseline and provides a lower bound on what can be achieved with adversarial networks or intensity-based data augmentation. As a second baseline, we (i) produce artificial CBCTs by passing CTs from cohort $\mathscr{C}_1$ through a cycleGAN [70] and (ii) use these artificial CBCTs to train a U-Net segmentation network. Code from Keras (https://keras.io/examples/generative/cyclegan/) has been adapted to train the 2D cycleGAN slice-wise on CTs from cohort $\mathscr{C}_1$ and CBCTs from cohort $\mathscr{C}_3$ for image translation. Finally, we compared our methods with PSIGAN [26], a UDA approach modeling the co-dependency between images and their segmentation as a joint probability distribution.

Registration baselines. In the context of CBCT segmentation for radiotherapy, the contoured planning CT scan is an important available prior. Hence, using deformable image registration (DIR), we can map the contours of the planning CT scan to the daily CBCT scan. Different mapping rules define different registration algorithms, and in this paper, three registration algorithms are compared: rigid, the registration method implemented in RayStation (a commercial treatment planning system), and Morphons. Rigid and RayStation are parametric methods: they look for the parameters of a deformations model – either linear (rigid) or non-linear (RayStation) – that yield the best match between the planning CT and daily CT scan. The diffeomorphic morphons' DIR algorithm, a non-parametric method, is implemented in OpenReggui (https://openreggui.org/) [25]. This method exploits the local phase of the scans to perform the registration. Therefore, it is suited for the registration of scans with different contrast enhancement, such as CT and CBCT scans. The diffeomorphic version of the algorithm forces anatomically plausible deformations. The registration has been performed between the CT and CBCT scans of patient cohort $\mathscr{C}_4$.

4. Results

In this section, we investigate three different strategies for improving cross-domain generalization of deep neural networks: adversarial networks, intensity-based data augmentation, and intensity-based data augmentation with a larger training set.

4.1. Adversarial networks

We train our adversarial networks with adaptations on L3, L6, L9, and L11, using the 30 CTs from cohort $\mathscr{C}_1$ and the 30 CBCTs from cohort $\mathscr{C}_3$. The CTs from cohort $\mathscr{C}_2$ are used to monitor the performance of the segmenter on source domain images. The performance of the generated models on CBCTs from $\mathscr{C}_4$ is compared with baseline methods (training on source and DIR algorithms) in Table 1. Compared with "source only" training, adversarial networks do not bring improvement when the adaptation is made on L3. An explanation is that since this layer corresponds to low-level features, discriminating between CT and CBCT does not make sense (detecting edges is relevant for both CT and CBCT). When the adaptation is performed on L6, a small improvement is observed for the bladder, while the performance for the prostate decreases slightly. When adapted on L9, the performance is similar for the bladder and prostate and slightly better for the rectum. Finally, when we adapt on L11 (i.e., the pre-softmax activations), the model performance is improved for all three organs, both in term of DSC and SMBD. This means that adversarial learning forces the features learned on CT to align with CBCTs so that they generalize to CBCT better. The adaptation proves especially useful for the rectum, which experiences the most dramatic change in intensities between the source and target domain images (see histograms in Fig. 3). Although slightly outperforming the rigid and RayStation (RS) intensity-based registration baselines (similar DSC and smaller SMBD) for the bladder, the performance of adversarial networks (L11) remains below these algorithms for the rectum and prostate, and below the morphons registration baseline for all three organs.

4.2. Intensity-based data augmentation

4.2.1. Comparison of different intensity-based data augmentation types

While adversarial networks align the features activated by source and target images, an alternative approach consists in altering the source images to align them with the target images. The performance of different intensity-based data augmentation strategies is compared in Table 2 using standard (non adversarial) training. In these experiments, we keep the same CTs (from cohort $\mathscr{C}_1$) as in the previous section for training and the same CBCTs (from cohort $\mathscr{C}_4$) for testing. Contrary to the previous section, images from cohorts $\mathscr{C}_2$ and $\mathscr{C}_3$ are not used. Strategies involving modification in levels of contrast, sharpness and noise are inefficient and perform similarly or worse than the "source only" baseline. Brightness, linear, and bilinear data augmentations, however, outperform the "source only" baseline on all three organs, in terms both of DSC and SMBD. Brightness and linear transformations perform similarly for all three organs, with slightly worse performance than morphons for the bladder, worse performance for the prostate, and better performance for the rectum. The bilinear method, finally, performs similarly to morphons for the bladder (larger DSC, larger SMBD) and worse on both the rectum and prostate. CycleGAN, when compared to the contrast, sharpness and noise transformations, performs better on some organs and worse on others. In contrary, the brightness, linear, and bilinear transformations all outperform cycleGAN. One possible explanation for the relative weakness of cycleGAN is that it generates blurry images. Blur is problematic because if decreases the contrast at organs boundaries, the most challenging areas for segmentation. An example of a slice of an original CT, as well as its transformation through cycleGAN and brightness operation is depicted in Fig. 4.

Since adversarial networks and brightness-based data augmentation both yield improvements over the baseline "source only" method, we investigate the performance of the combined approaches. We chose brightness over other intensity-based data augmentation strategies because it has the best balance of performance over the three organs. Combining both approaches led to no improvement compared with brightness augmentation alone (see last raw in Table 2).

Table 1
Performance (mean±std) on the 30 CBCTs from cohort $\mathscr{C}_4$ for adversarial networks and for the baseline methods (train on source and DIRs). DSC: Dice similarity coefficient, SMBD: symmetric mean boundary distance, DL: deep learning, DIR: deformable image registration, RS: RayStation. ◊ For at least one patient, the model predicts no segmentation at all for the organ and therefore the SMBD of that patient (as well as the mean over all patients) is not defined.

Method	DSC			SMBD (mm)		
	Bladder	Rectum	Prostate	Bladder	Rectum	Prostate
DIR, Rigid image registration	0.757±0.150	0.629±0.087	0.721±0.130	5.48±3.53	**5.77±1.96**	3.90±1.67
DIR, RS intensity-based	0.775±0.154	0.640±0.096	0.720±0.133	5.01±3.50	5.61±2.12	3.91±1.74
DIR, Morphons	**0.813±0.154**	**0.653±0.169**	**0.731±0.146**	**4.00±3.17**	5.78±4.08	**3.63±1.83**
DL, Train on source	0.749±0.212	0.179±0.241	0.629±0.169	5.76±5.96	Undefined◊	4.96±2.29
DL, Adversarial (L3)	0.764±0.180	0.168±0.243	0.636±0.137	5.63±5.21	Undefined◊	5.04±2.19
DL, Adversarial (L6)	0.771±0.164	0.193±0.263	0.609±0.145	4.88±3.48	Undefined◊	5.41±2.08
DL, Adversarial (L9)	0.735±0.192	0.210±0.261	0.624±0.158	4.98±3.50	Undefined◊	5.26±2.42
DL, Adversarial (L11)	0.787±0.166	0.447±0.181	0.660±0.158	4.60±3.74	8.04±3.13	4.72±2.40

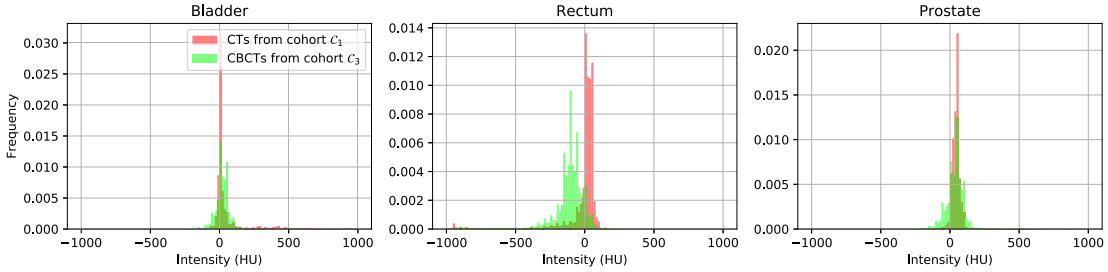
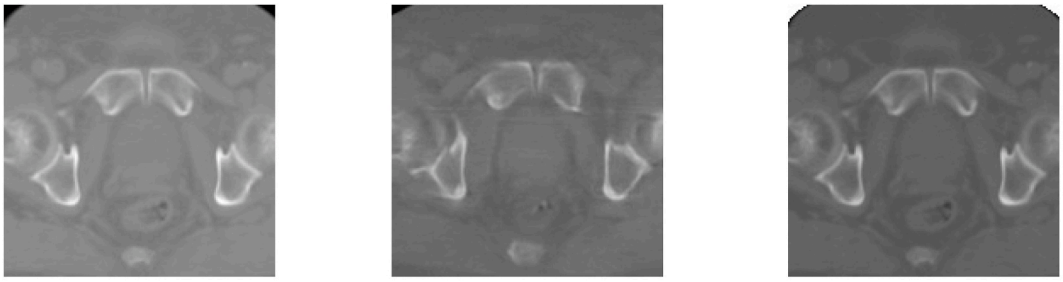


Fig. 3. Histograms differ significantly between the 30 CTs from cohort $\mathscr{C}_1$ and the 30 cone beam CTs from cohort $\mathscr{C}_3$ (the two cohorts used for training the adversarial networks), especially inside the rectum.

Table 2
Performance (mean±std) on the 30 CBCTs from cohort $\mathscr{C}_4$ for intensity-based data augmentation and for the baseline methods (training on source and DIRs). DSC: dice similarity coefficient, SMBD: symmetric mean boundary distance, DL: deep learning, DIR: deformable image registration, RS: RayStation, Adv.: adversarial networks. ◊ For at least one patient, the model predicts no segmentation at all for the organ and therefore the SMBD of that patient (as well as the mean over all patients) is not defined.

Method	DSC			SMBD (mm)		
	Bladder	Rectum	Prostate	Bladder	Rectum	Prostate
DIR, Morphons	0.813±0.154	0.653±0.169	**0.731±0.146**	**4.00±3.17**	5.78±4.08	**3.63±1.83**
DL, Train on source	0.749±0.212	0.179±0.241	0.629±0.169	5.76±5.96	Undefined◊	4.96±2.29
DL, CycleGAN	0.744±0.140	0.602±0.119	0.521±0.146	5.77±3.39	5.69±2.13	6.86±2.32
DL, Brightness	0.791±0.191	0.690±0.121	0.658±0.135	4.38±4.08	**4.78±2.03**	5.08±3.08
DL, Contrast	0.765±0.179	0.205±0.247	0.618±0.180	6.46±5.20	Undefined◊	5.47±2.96
DL, Sharpness	0.661±0.164	0.183±0.252	0.584±0.182	6.48±3.04	Undefined◊	5.70±2.51
DL, Noise	0.672±0.173	0.158±0.234	0.534±0.199	6.91±4.21	35.06±26.99	8.96±5.39
DL, Linear	0.775±0.125	**0.694±0.100**	0.651±0.133	4.70±2.85	5.11±2.37	5.07±2.63
DL, Bilinear	**0.837±0.107**	0.562±0.174	0.666±0.117	4.67±3.60	6.13±2.67	4.77±2.42
DL, Brightness + adv.	0.777±0.157	**0.694±0.097**	0.654±0.137	4.83±3.72	**4.78±1.92**	5.08±2.52



(a) Original CT. (b) CT transformed with cycleGAN. (c) CT transformed with brightness.

Fig. 4. Example of intensity-based data augmentations.

4.2.2. Brightness with the extended source dataset

In Section 4.2, we used only 30 training CTs (from cohort $\mathscr{C}_1$) in order to make a fair comparison with adversarial networks. As a last experiment, we investigated whether using all available CTs (74 examples for cohorts $\mathscr{C}_1$, $\mathscr{C}_2$, and $\mathscr{C}_5$) improved the performance. Table 3 shows that the proposed method performs as well as morphons for the prostate, and outperforms it on the bladder and rectum. Actual segmentations on a representative patient are shown in Fig. 5. It also outperforms adversarial networks (L11) on all three organs.

5. Discussion

U-Net trained on CTs for bladder, rectum and prostate segmentation does not generalize well on CBCTs, as shown in Table 3. For the rectum, in particular, it produces many false negatives (this is illustrated in Fig. 5). This may be explained by the fact that deep learning learns correlations between patterns of intensities and presence/absence of a given organ, without prior information about anatomy. For instance, a network trained in a non-adversarial way has no incentive to avoid giving a zero prediction for the rectum, even if every patient has one. Image registration methods, in contrast, deform the anatomy from planning to the daily image. We show that adversarial networks applying adaptation of the pre-softmax activations (L11) yield minor improvements (compared to source-only training) on the bladder and prostate, and significant improvement on the rectum (for this last organ, the DSC more than doubles). For this organ, the improvement can be attributed to the introduction of an anatomical a priori. By encouraging the network to have indistinguishable pre-softmax activations between CT and CBCT, UDA forces the network to produce contours for the rectum, even if the intensities in the test CBCTs are not similar to the ones observed in the training CTs.

In the context of radiotherapy, one disposes of an interesting a priori information to segment the CBCT, namely, the planning CT. Registration takes advantage of this information. This algorithm is a relevant baseline because, similar to the two methods proposed in this paper (adversarial networks and intensity-based data augmentation), it does not have to be trained on annotated CBCTs. Adversarial networks' performance compared to registration's is contrasted. We applied three registration algorithms to our dataset. Adversarial networks perform better than rigid registration and similar to the RayStation commercial software on the bladder and perform worse on the two other organs. Compared with the morphons algorithm, adversarial networks (irrespective of the layer adapted) perform worse on all three organs. We also compared our results to other works on other datasets. For the bladder, the comparison is mixed, with adversarial networks (L11) performing worse [31,62], similarly [42] or better [55,57], depending on the study. For the rectum and prostate, adversarial networks (L11) perform worse than most other authors' methods, with the exception of MIM intensity-based [42]. Note that those studies report results on a dataset different from ours. It is therefore difficult to determine what part of the difference in performance can be attributed to the method used, and what part to differences in the degree of difficulty in the test datasets. A limitation of the comparison shown in Table 3 is that it does not compare our methods with other unsupervised domain adaptation frameworks on our data.

An advantage of adversarial networks over registration is faster inference. For each new test image, registration requires iterative image alignment with gradient descent, which takes several minutes. This prevents its application in the context of adaptive treatment, when the patient lies on the couch while treatment adaptation is being computed. In contrast, while adversarial networks require long (a few hours) training times, prediction does not entail iterative optimization but only matrix multiplication, which takes less than a second.

For the bladder, we compared adversarial networks with two patient-specific models. Here the conclusion is also mitigated, with adversarial networks (L11) performing worse than one [8] and better than the other [51]. A drawback of patient specific-models is that they require contouring several images of the target patient by hand, which is long, tedious, and therefore often not practical. This is not necessary for adversarial networks.

While the performance of adversarial networks compared with

Table 3
Performance (mean±std) on the 30 CBCTs from cohort $\mathscr{C}_4$ for adversarial networks, intensity-based data augmentation, baseline methods as well as the performance of previous works on other datasets. $^{\diamondsuit}$ For at least one patient, the model predicts no segmentation at all for the organ and therefore the SMBD of that patient (as well as the mean over all patients) is not defined. † Results reported on a test set containing both CBCT and CT scans. ‡ The authors computed the root mean squared boundary distance rather than the SMBD. § The authors computed the mean boundary distance and not the SMBD. For deep learning methods, n_{CBCT} and n_{CT} are the number of annotated CBCTs and CTs used for training (respectively).

Study	Method	DSC			SMBD (mm)		
		Bladder	Rectum	Prostate	Bladder	Rectum	Prostate
Ours	DIR, Morphons	0.813±0.154	0.653±0.169	0.731±0.146	4.00±3.17	5.78±4.08	**3.63±1.83**
Ours	DL, Train on source ($n_{\text{CBCT}}=0, n_{\text{CT}}=30$)	0.749±0.212	0.179±0.241	0.629±0.169	5.76±5.96	Undefined$^{\diamondsuit}$	4.96±2.29
Ours	DL, Adversarial (L11) ($n_{\text{CBCT}}=0, n_{\text{CT}}=30$)	0.787±0.166	0.447±0.181	0.660±0.158	4.60±3.74	8.04±3.13	4.72±2.40
Ours	DL, Train on source ($n_{\text{CBCT}}=0, n_{\text{CT}}=74$)	0.775±0.138	0.173±0.235	0.628±0.145	5.11±3.25	Undefined$^{\diamondsuit}$	6.65±3.78
Ours	DL, Brightness ($n_{\text{CBCT}}=0, n_{\text{CT}}=74$)	0.837±0.142	0.701±0.114	**0.734±0.113**	3.55±2.44	4.80±2.13	**3.63±1.79**
Ours	DL, PSIGAN ($n_{\text{CBCT}}=0, n_{\text{CT}}=30$)	0.825±0.113	0.747±0.083	0.674±0.102	3.74±2.11	3.10±1.19	4.16±1.47
Ours	DL, PSIGAN ($n_{\text{CBCT}}=0, n_{\text{CT}}=74$)	**0.868±0.094**	**0.762±0.093**	0.719±0.083	**2.90±2.09**	**2.96±1.23**	3.66±1.36
Ours	DL, PSIGAN + Brightness ($n_{\text{CBCT}}=0, n_{\text{CT}}=74$)	0.850±0.110	0.697±0.140	0.685±0.101	3.98±2.86	3.62±1.69	3.90±1.66
[17]	DL ($n_{\text{CBCT}}=80, n_{\text{CT}}=0$)	**0.96±0.03**	**0.93±0.04**	**0.91±0.08**	**0.65±0.67**§	**0.72±0.61**§	**0.93±0.96**§
[32]	DL ($n_{\text{CBCT}}=42, n_{\text{CT}}=74$)	0.874±0.096	0.814±0.055	0.758±0.101	2.47±1.93	2.38±0.98	3.08±1.48
[52]	DL ($n_{\text{CBCT}}=300$, $n_{\text{CT}}=300$)	0.932†	0.871†	0.840†	2.57 ± 0.54†,‡	2.47 ± 0.64†,‡	2.34 ± 0.68†,‡
[21]	DL ($n_{\text{CBCT}}=124$, $n_{\text{CT}}=88$)	0.88	0.71	–	–	–	–
[42]	DIR, MIM intensity-based	~ 0.80	~ 0.40	~ 0.55	–	–	–
[42]	DIR, RS intensity-based	~ 0.78	~ 0.70	~ 0.75	–	–	–
[55]	DIR, RS intensity-based	0.69 ± 0.07	0.75 ± 0.05	0.84 ± 0.05	–	–	–
[62]	DIR, Cascade MI-based	~ 0.83	~ 0.77	~ 0.80	~ 2.6§	~ 2.3§	~ 2.3§
[31]	DIR, Rigid on bone and prostate	0.85 ± 0.05	–	0.82 ± 0.04	–	–	–
[57]	DIR, Demons	0.73	0.77	0.80			
[51]	Patient specific model	~ 0.87	–	–	–	–	–
[8]	Patient specific model	0.78	–	–	–	–	–

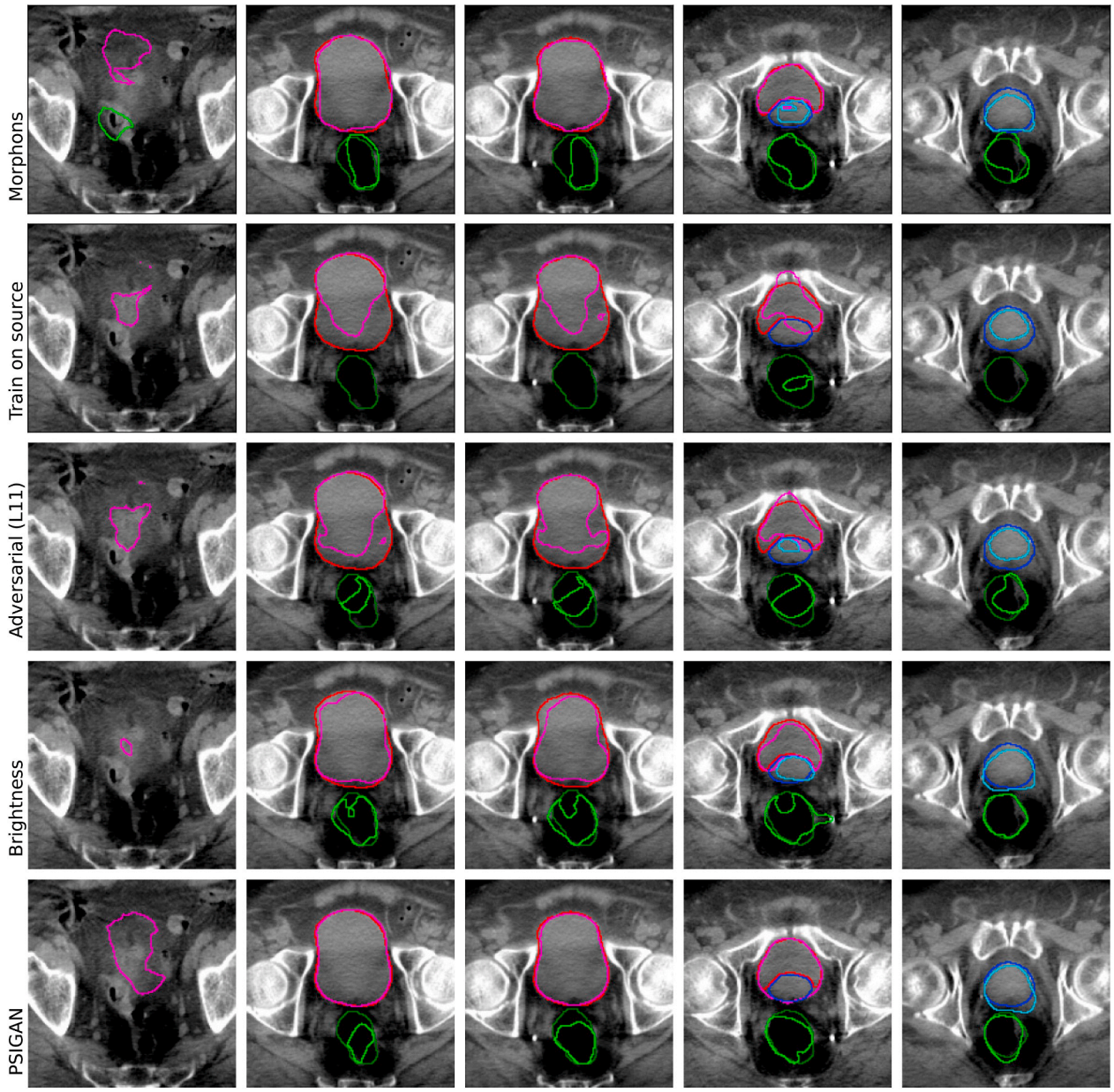


Fig. 5. Comparison of morphons, "train on source", adversarial networks with adaptation on L11, brightness, and PSIGAN segmentation for a representative patient from cohort $\mathscr{C}_4$. Each column corresponds to a slice of the same CBCT scan. Dark colors represent reference segmentations, while light colors show predictions from the different algorithms. Compared to "train on source", Adversarial networks (L11) brings small improvements for the bladder (by detecting slightly larger portions of it) and the prostate (it is missed by "train on source" in the fourth column), as well as large improvement in rectum (since it is completely missed by "source on train" on most slices). Brightness performs similar to morphons for the prostate, slightly outperforms it for the bladder (fewer false negatives in the first column) and performs better for the rectum (no false negatives in the first column and more accurate delineation in the fifth). Color should be used for this figure in print.

registration and patient-specific models is contrasted, this method is interesting since, to our knowledge, this is the first time that UDA has been applied to segmentation using U-Net with two settings: long skip connections and fully 3D. Previous work chose to remove the long skip connections when performing UDA, probably out of the concern that the network could use them to "bypass" the features alignment constraint. Analyzing the impact of skip connections for adversarial networks in UDA settings remains to be addressed and would be interesting to investigate in future studies. A second interesting feature of our approach is that it is built on a 3D U-Net instead of aggregating the outputs from a 2D U-Net, which is what is most often done with UDA in segmentation. Using 3D leverages useful information from an additional dimension. A possible worry about using 3D instead of 2D is that it artificially reduces the number of different samples seen by the discriminator (by a factor equal to the number of slices per image). Again, although the comparison between 2D and 3D is beyond the scope of this study, we have shown that training a 3D network is possible, even if few samples are available for training.

A challenge with adversarial networks is that their training is notoriously unstable. In this paper, we also analyzed an alternative method where input images are transformed so that they look more like CBCTs. The method with best average performance across organs was brightness-based augmentation, which consists in adding a random offset on the training images. For fair comparison with adversarial networks, a model with brightness augmentation was first trained on 30 CTs, performing similarly to adversarial networks (L11) for the prostate and better on the bladder and rectum. When training is performed on a larger set of 74 CTs, this method performs similarly to morphons on

prostate and better on the two other organs. Compared with previous studies based on DIR, brightness with the extended dataset performs generally similarly or better on the bladder. For the rectum and the prostate, however, our algorithm performs equally well or worse (it is better than MIM intensity-based from Ref. [42] only). To summarize this comparison between intensity-based data augmentation and previous work using other strategies, our method thus has relevance for the bladder, as well as for applications where short inference time matters, such as adaptive radiotherapy.

Table 3 summarizes our better results, and presents the results obtained from DIR methods, as well as DL methods trained fully on CBCT data or a combination of both CT and CBCT. Our results cannot be directly compared with these DL methods due to the inherent challenge of training with unlabeled data, which is the bottleneck of current clinical practice. In addition, the number of patients used in these studies is often larger than our database (Schreier et al., for example, used 300 patients [52]). Acquiring databases of patients larger than 50–100 also represents an added complexity for small to medium-sized hospitals. Given the small database and the challenge of training with unlabeled CBCTs, our method achieves very good results: a max. difference of only 0.1–0.2 in the dice for the best reported results in the literature. In addition, the results in Section 4.3 suggest that increasing the size of the database for adversarial training, too, may lead to an improvement of the results. The PSIGAN method with 74 training CTs outperforms the brightness-based data augmentation on the bladder and rectum but performs worse on the prostate. This suggests that modeling the joint distribution of images and their segmentation can be useful for CT to CBCT adaptation. We investigated adding brightness-based data augmentation on top of PSIGAN, but it hurt performance. Similarly, we saw in Section 4.2.1 that adding brightness-based data augmentation on top of gradient reversal layer networks decreased performance for the bladder and prostate, even it brought improvement on the rectum. Eventually, brightness-based data augmentation is an interesting technique that is simple and works well alone, but with ambiguous effects when combined with other UDA methods.

We did not use batch normalization since it hurted performance. A similar situation was observed in the original 3D U-Net paper [13] for fully-automated segmentation. A possible explanation is that there is too little variety in the training batches with our training set of only 30 images. For this reason, methods relying on batch normalization for domain adaptation [35] cannot be used in our situation.

6. Conclusions

In this study, we propose an extension of 3D U-Net on unsupervised domain adaptation, applying adversarial networks to the segmentation of organs (bladder, rectum and prostate) in cone beam CT. For a model trained on CT, this method improves generalization to CBCTs, using only non-annotated CBCTs. Our code is made public and can be used for any application experiencing a shift of distribution between training and test images, a recurring problem in biomedical imaging. Although less accurate than morphons, a registration-based algorithm, this method is faster (less than one second versus several minutes). This is essential for online treatment adaptation. We also show that when intensity-based data augmentations are applied to training CTs, the network generalizes better to CBCT, reaching a level similar to (prostate) or better (bladder, rectum) than the best registration-based algorithms. Intensity-based data augmentation is best used alone since it has ambiguous effects when combined with other UDA methods. In the future, this could be used in radiotherapy treatment to measure and ultimately reduce the dose delivered to healthy organs, while delivering the prescribed dose to the target.

Funding

Eliott Brion is supported by the Walloon Region under Grant RW-DGO6-Biowin-Bidmed. Jean Léger is a Research Fellow of the Belgian national scientific research foundation Fonds de la Recherche Scientifique (FNRS). Ana Barragán is funded by the Walloon region (PROTHERWAL/CHARP, grant number 7289). John A. Lee is a Senior Research Associates with the Belgian F.R.S.-FNRS. The funding sources have no involvement in study design; in the collection, analysis and interpretation of data; in the writing of the report; and in the decision to submit the article for publication.

Declaration of competing interest

The authors have no competing interests to declare.

Acknowledgments

We thank CHU-UCL-Namur and CHU-Charleroi for providing the data. The authors thank Sara Teruel Rivas for her technical support in the data acquisition and annotation and to Dr. Geneviève Van Ooteghem for verification of the contours. Finally, we thank Gabrielle Leyden for editing the final revision of this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2021.104269.

CRediT authorship contribution statement

Eliott Brion: Conceptualization, Methodology, Software, Data Curation, Writing - Original Draft. Jean Léger: Conceptualization, Methodology, Data Curation, Writing - Original Draft. A.M. Barragán-Montero: Conceptualization, Methodology, Writing - Review Editing. Nicolas Meert: Data Curation, Writing - Review Editing. John A. Lee: Conceptualization, Methodology, Writing - Review Editing, Supervision, Funding acquisition. Benoit Macq: Conceptualization, Methodology, Writing - Review Editing, Supervision, Funding acquisition.

References

[1] H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, Domain-adversarial Neural Networks, 2014 arXiv:1412.4446.
[2] K. Armanious, C. Jiang, M. Fischer, T. Küstner, T. Hepp, K. Nikolaou, S. Gatidis, B. Yang, Medgan: medical image translation using gans, Comput. Med. Imag. Graph. 79 (2020) 101684.
[3] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, D. Krishnan, Unsupervised pixel-level domain adaptation with generative adversarial networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3722–3731.
[4] C. Boydev, D. Pasquier, F. Derraz, L. Peyrodie, A. Taleb-Ahmed, J.P. Thiran, Automatic prostate segmentation in cone-beam computed tomography images using rigid registration, in: 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2013, pp. 3993–3997.
[5] E. Brion, J. Léger, U. Javaid, J. Lee, C. De Vleeschouwer, B. Macq, Using planning cts to enhance cnn-based bladder segmentation on cone beam ct, in: Medical Imaging 2019: Image-Guided Procedures, Robotic Interventions, and Modeling, International Society for Optics and Photonics, 2019, p. 109511M.
[6] J. Cai, Z. Zhang, L. Cui, Y. Zheng, L. Yang, Towards cross-modal organ translation and segmentation: a cycle-and shape-consistent generative adversarial network, Med. Image Anal. 52 (2019) 174–184.
[7] K.H. Cha, L. Hadjiiski, R.K. Samala, H.P. Chan, E.M. Caoili, R.H. Cohan, Urinary bladder segmentation in ct urography using deep-learning convolutional neural network and level sets, Med. Phys. 43 (2016) 1882–1896.
[8] X. Chai, M. van Herk, A. Betgen, M. Hulshof, A. Bel, Automatic bladder segmentation on cbct for multiple plan art of bladder cancer using a patient-specific bladder model, Phys. Med. Biol. 57 (2012) 3945.
[9] C. Chen, Q. Dou, H. Chen, P.A. Heng, Semantic-aware generative adversarial nets for unsupervised domain adaptation in chest x-ray segmentation, in: International Workshop on Machine Learning in Medical Imaging, Springer, 2018, pp. 143–151.
[10] C. Chen, Q. Dou, H. Chen, J. Qin, P.A. Heng, Synergistic image and feature adaptation: towards cross-modality domain adaptation for medical image segmentation, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2019, pp. 865–872.

[11] V. Cheplygina, M. de Bruijne, J.P. Pluim, Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis, Med. Image Anal. 54 (2019) 280–296.

[12] J. Choi, T. Kim, C. Kim, Self-ensembling with gan-based data augmentation for domain adaptation in semantic segmentation, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 6830–6840.

[13] Ö. Çiçek, A. Abdulkadir, S.S. Lienkamp, T. Brox, O. Ronneberger, 3d u-net: learning dense volumetric segmentation from sparse annotation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2016, pp. 424–432.

[14] J.P. Cohen, M. Luck, S. Honari, Distribution matching losses can hallucinate features in medical image translation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2018, pp. 529–536.

[15] Q. Dou, C. Ouyang, C. Chen, H. Chen, P.A. Heng, Unsupervised Cross-Modality Domain Adaptation of Convnets for Biomedical Image Segmentations with Adversarial Loss, 2018 arXiv preprint arXiv:1804.10916.

[16] G. French, M. Mackiewicz, M. Fisher, Self-ensembling for Visual Domain Adaptation, 2017 arXiv preprint arXiv:1706.05208.

[17] Y. Fu, Y. Lei, T. Wang, S. Tian, P. Patel, A.B. Jani, W.J. Curran, T. Liu, X. Yang, Pelvic multi-organ segmentation on cone-beam ct for prostate adaptive radiotherapy, Med. Phys. 47 (8) (2020) 3415–3422.

[18] Y. Ganin, V. Lempitsky, Unsupervised Domain Adaptation by Backpropagation, 2014 arXiv preprint arXiv:1409.7495.

[19] M. Ghafoorian, A. Mehrtash, T. Kapur, N. Karssemeijer, E. Marchiori, M. Pesteie, C. R. Guttmann, F.E. de Leeuw, C.M. Tempany, B. Van Ginneken, et al., Transfer learning for domain adaptation in mri: application in brain lesion segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2017, pp. 516–524.

[20] M. Ghilezan, D. Yan, A. Martinez, Adaptive radiation therapy for prostate cancer, Semin. Radiat. Oncol. 20 (2010) 130–137, https://doi.org/10.1016/j.semradonc.2009.11.007. URL:.

[21] A. Haensch, V. Dicken, T. Gass, T. Morgas, J. Klein, H. Meine, H. Hahn, Deep learning based segmentation of organs of the female pelvis in cbct scans for adaptive radiotherapy using ct and cbct data, Int. J. Comput. Assist. Radiol. Surg. 13 (2018) 179–180.

[22] A. Hänsch, V. Dicken, J. Klein, T. Morgas, B. Haas, H.K. Hahn, Artifact-driven sampling schemes for robust female pelvis cbct segmentation using deep learning, in: Medical Imaging 2019: Computer-Aided Diagnosis, International Society for Optics and Photonics, 2019, p. 109500T.

[23] J. Hoffman, E. Tzeng, T. Park, J.Y. Zhu, P. Isola, K. Saenko, A. Efros, T. Darrell, Cycada: cycle-consistent adversarial domain adaptation, in: International Conference on Machine Learning, 2018, pp. 1989–1998.

[24] Y. Huo, Z. Xu, H. Moon, S. Bao, A. Assad, T.K. Moyo, M.R. Savona, R.G. Abramson, B.A. Landman, Synseg-net: synthetic segmentation without target modality ground truth, IEEE Trans. Med. Imag. 38 (2018) 1016–1025.

[25] G. Janssens, L. Jacques, J. Orban de Xivry, X. Geets, B. Macq, Diffeomorphic registration of images with variable contrast enhancement, Int. J. Biomed. Imag. (2011), 2011.

[26] J. Jiang, Y.C. Hu, N. Tyagi, A. Rimner, N. Lee, J.O. Deasy, S. Berry, H. Veeraraghavan, Psigan: joint probabilistic segmentation and image distribution matching for unpaired cross-modality adaptation based mri segmentation, IEEE Trans. Med. Imag. (2020).

[27] J. Jiang, Y.C. Hu, N. Tyagi, P. Zhang, A. Rimner, G.S. Mageras, J.O. Deasy, H. Veeraraghavan, Tumor-aware, adversarial domain adaptation from ct to mri for lung cancer segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2018, pp. 777–785.

[28] K. Kamnitsas, C. Baumgartner, C. Ledig, V. Newcombe, J. Simpson, A. Kane, D. Menon, A. Nori, A. Criminisi, D. Rueckert, et al., Unsupervised domain adaptation in brain lesion segmentation with adversarial networks, in: International Conference on Information Processing in Medical Imaging, Springer, 2017, pp. 597–609.

[29] S. Kazemifar, A. Balagopal, D. Nguyen, S. McGuire, R. Hannan, S. Jiang, A. Owrangi, Segmentation of the Prostate and Organs at Risk in Male Pelvic CT Images Using Deep Learning, 2018 arXiv preprint arXiv:1802.09587.

[30] M. Kim, H. Byun, Learning texture invariant representation for domain adaptation of semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 12975–12984.

[31] L. König, A. Derksen, N. Papenberg, B. Haas, Deformable image registration for adaptive radiotherapy with guaranteed local rigidity constraints, Radiat. Oncol. 11 (2016) 122.

[32] J. Léger, E. Brion, P. Desbordes, C. De Vleeschouwer, J.A. Lee, B. Macq, Cross-domain data augmentation for deep-learning-based male pelvic organ segmentation in cone beam ct, Appl. Sci. 10 (2020) 1154.

[33] Y. Lei, T. Wang, S. Tian, X. Dong, A.B. Jani, D. Schuster, W.J. Curran, P. Patel, T. Liu, X. Yang, Male pelvic multi-organ segmentation aided by cbct-based synthetic mri, Phys. Med. Biol. 65 (2020), 035013.

[34] H. Li, T. Loehr, B. Wiestler, J. Zhang, B. Menze, E-Uda: Efficient Unsupervised Domain Adaptation for Cross-Site Medical Image Segmentation, 2020 arXiv preprint arXiv:2001.09313.

[35] Y. Li, N. Wang, J. Shi, J. Liu, X. Hou, Revisiting Batch Normalization for Practical Domain Adaptation, 2016 arXiv preprint arXiv:1603.04779.

[36] Y. Li, L. Yuan, N. Vasconcelos, Bidirectional learning for domain adaptation of semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 6936–6945.

[37] F. Liu, Susan: segment unannotated image structure using adversarial network, Magn. Reson. Med. 81 (2019) 3330–3345.

[38] M. Long, Y. Cao, J. Wang, M. Jordan, Learning transferable features with deep adaptation networks, in: International Conference on Machine Learning, 2015, pp. 97–105.

[39] Y. Luo, P. Liu, T. Guan, J. Yu, Y. Yang, Significance-aware information bottleneck for domain adaptive semantic segmentation, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 6778–6787.

[40] R. Ma, P. Tao, H. Tang, Optimizing data augmentation for semantic segmentation on small-scale dataset, in: Proceedings of the 2nd International Conference on Control and Computer Vision, 2019, pp. 77–81.

[41] A. Mikołajczyk, M. Grochowski, Data augmentation for improving deep learning in image classification problem, in: 2018 International Interdisciplinary PhD Workshop, IIPhDW, IEEE, 2018, pp. 117–122.

[42] K. Motegi, H. Tachibana, A. Motegi, K. Hotta, H. Baba, T. Akimoto, Usefulness of hybrid deformable image registration algorithms in prostate radiation therapy, J. Appl. Clin. Med. Phys. 20 (2019) 229–236.

[43] J. Nalepa, M. Marcinkiewicz, M. Kawulok, Data augmentation for brain-tumor segmentation: a review, Front. Comput. Neurosci. 13 (2019).

[44] M. Nishio, S. Noguchi, K. Fujimoto, Automatic pancreas segmentation using coarse-scaled 2d model of deep learning: usefulness of data augmentation and deep u-net, Appl. Sci. 10 (2020) 3360.

[45] P. Novosad, V. Fonov, D.L. Collins, Unsupervised Domain Adaptation for the Automated Segmentation of Neuroanatomy in Mri: a Deep Learning Approach, bioRxiv, 2019, p. 845537.

[46] S. Oh, S. Kim, Deformable image registration in radiation therapy, Radiat. Oncol. J. 35 (2017) 101.

[47] C. Peng, E. Ahunbay, G. Chen, S. Anderson, C. Lawton, X.A. Li, Characterizing interfraction variations and their dosimetric effects in prostate cancer radiotherapy, Int. J. Radiat. Oncol. Biol. Phys. 79 (2011) 909–914.

[48] C.S. Perone, P. Ballester, R.C. Barros, J. Cohen-Adad, Unsupervised domain adaptation for medical imaging segmentation with self-ensembling, Neuroimage 194 (2019) 1–11.

[49] F. Pos, P. Remeijer, Adaptive management of bladder cancer radiotherapy, in: Seminars in Radiation Oncology, Elsevier, 2010, pp. 116–120.

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

[51] A. van de Schoot, G. Schooneveldt, S. Wognum, M. Hoogeman, X. Chai, L. Stalpers, C. Rasch, A. Bel, Generic method for automatic bladder segmentation on cone beam ct using a patient-specific bladder shape model, Med. Phys. 41 (2014).

[52] J. Schreier, A. Genghi, H. Laaksonen, T. Morgas, B. Haas, Clinical evaluation of a full-image deep segmentation algorithm for the male pelvis on cone-beam ct and ct, Radiother. Oncol. 145 (2020) 1–6.

[53] C. Shorten, T.M. Khoshgoftaar, A survey on image data augmentation for deep learning, J. Big Data 6 (2019) 60.

[54] B. Sun, J. Feng, K. Saenko, Correlation alignment for unsupervised domain adaptation, in: Domain Adaptation in Computer Vision Applications, Springer, 2017, pp. 153–171.

[55] Y. Takayama, N. Kadoya, T. Yamamoto, K. Ito, M. Chiba, K. Fujiwara, Y. Miyasaka, S. Dobashi, K. Sato, K. Takeda, et al., Evaluation of the performance of deformable image registration between planning ct and cbct images for the pelvic region: comparison between hybrid and intensity-based dir, J. Radiat. Res. 58 (2017) 567–571.

[56] A. Tarvainen, H. Valpola, Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results, Adv. Neural Inf. Process. Syst. (2017) 1195–1204.

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

[58] E. Tzeng, J. Hoffman, T. Darrell, K. Saenko, Simultaneous deep transfer across domains and tasks, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 4068–4076.

[59] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, T. Darrell, Deep Domain Confusion: Maximizing for Domain Invariance, 2014 arXiv preprint arXiv:1412.3474.

[60] V. Varkarakis, S. Bazrafkan, P. Corcoran, Deep neural network and data augmentation methodology for off-axis iris segmentation in wearable headsets, Neural Network. 121 (2020) 101–121.

[61] O. Weistrand, S. Svensson, The anaconda algorithm for deformable image registration in radiotherapy, Med. Phys. 42 (2015) 40–53.

[62] A.J. Woerner, M. Choi, M.M. Harkenrider, J.C. Roeske, M. Surucu, Evaluation of deformable image registration-based contour propagation from planning ct to cone-beam ct, Technol. Canc. Res. Treat. 16 (2017) 801–810.

[63] Y. Xia, D. Yang, Z. Yu, F. Liu, J. Cai, L. Yu, Z. Zhu, D. Xu, A. Yuille, H. Roth, Uncertainty-aware multi-view co-training for semi-supervised medical image segmentation and domain adaptation, Med. Image Anal. (2020) 101766.

[64] J. Yang, N.C. Dvornek, F. Zhang, J. Chapiro, M. Lin, J.S. Duncan, Unsupervised domain adaptation via disentangled representations: application to cross-modality liver segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2019, pp. 255–263.

[65] V. Zambrano, H. Furtado, D. Fabri, C. Lütgendorf-Caucig, J. Góra, M. Stock, R. Mayer, W. Birkfellner, D. Georg, Performance validation of deformable image registration in the pelvic region, J. Radiat. Res. 54 (2013) i120–i128.

[66] Y. Zhang, P. David, B. Gong, Curriculum domain adaptation for semantic segmentation of urban scenes, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2020–2030.

[67] Z. Zhang, L. Yang, Y. Zheng, Translating and segmenting multimodal medical volumes with cycle-and shape-consistency generative adversarial network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 9242–9251.
[68] H. Zhao, H. Li, S. Maurer-Stroh, Y. Guo, Q. Deng, L. Cheng, Supervised segmentation of un-annotated retinal fundus images by synthesis, IEEE Trans. Med. Imag. 38 (2018) 46–56.
[69] J.Y. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Computer Vision (ICCV), 2017 IEEE International Conference on, 2017.
[70] J.Y. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2223–2232.
[71] Y. Zou, Z. Yu, B. Vijaya Kumar, J. Wang, Unsupervised domain adaptation for semantic segmentation via class-balanced self-training, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 289–305.