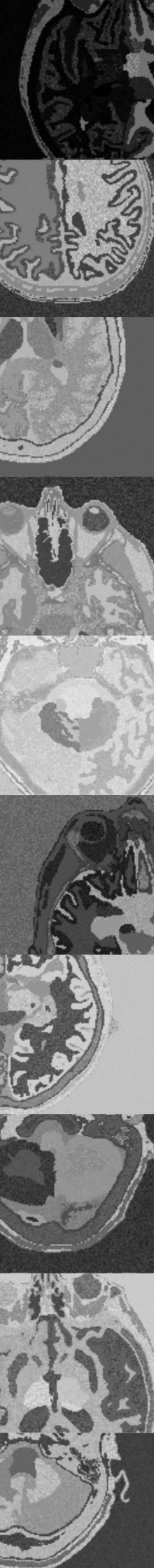

UMC Utrecht

UNIVERSITY OF TWENTE.

Faculty of Science and Technology (TNW)
Magnetic Detection & Imaging (MDI)

Contrast generalisation in deep learning-based brain MRI-to-CT synthesis

M.Sc. Assignment, Biomedical Engineering
Track: Imaging & In Vitro Diagnostics

L. Nijskens
June 20, 2022

Graduation committee:
prof. dr. ir. B. ten Haken, University of Twente
dr. ir. F.F.J. Simonis, University of Twente
dr. F.J. Siepel, University of Twente
prof. dr. ir. C.A.T. van den Berg, UMC Utrecht
dr. M. Maspero, UMC Utrecht

Faculty of Science and Technology (TNW),
Magnetic Detection & Imaging (MDI)
University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands

Samenvatting (Nederlands)

Achtergrond en doel: Computertomografie (CT) is de basis voor radiotherapie (RT) planning en bevat de voor dosisberekeningen benodigde informatie over elektronendichtheid. Magnetic resonance imaging (MRI) biedt superieur contrast in de weke delen en kan helpen bij het intekenen van tumoren, maar bevat geen inherente informatie over elektronendichtheid, in tegenstelling tot CT. Derhalve worden CT en MRI vaak gecombineerd bij het maken van een bestralingsplan, wat vraagt om beeldregistratie. RT op basis van alleen MRI is voorgesteld om overgebleven fouten na registratie te vermijden, blootstelling van de patiënt aan ioniserende straling te verminderen en de workflow te versimpelen. Het maken van synthetische CT (sCT) is noodzakelijk voor MRI-geleide RT planning, iets wat binnen seconden kan met deep learning (DL). De acquisitieprotocollen voor MRI kunnen veranderen over de tijd of verschillen tussen zorginstellingen. Zonder her-training generaliseren DL modellen slecht naar nieuwe domeinen, zoals verschillende acquisitieprotocollen. Dit verhindert hun wijdverbreide gebruik. *Domain randomisation* is een leermethode die gebaseerd is op het genereren van trainingsdata met gerandomiseerde parameters. De methode heeft tot veelbelovende resultaten geleid voor het verbeteren van generaliseerbaarheid, bijvoorbeeld voor segmentatie. Dit werk onderzoekt het vermogen van DL modellen voor het genereren van sCT van de hersenen om te generaliseren naar MRI scans gemaakt met een ongeziene sequentie zonder het network te her-trainen, alsook hoe domain randomisation de netwerkprestaties op deze ongeziene sequentie beïnvloedt.

Materialen and methoden: Data van 95 patiënten die RT hebben ondergaan werden geïncludeerd uit een retrospectieve database. Hierbij waren een CT met bijbehorende T_1-gewogen MRI met en zonder contrastmiddel (T1wGd en T1w), T_2-gewogen (T2w) en FLAIR MRI vereist voor iedere patiënt. Een Baseline *conditional generative adversarial network* werd getraind met en zonder ongeziene sequentie (FLAIR) om te testen hoe een netwerk presteert op een ongeziene sequentie zonder domain randomisation. Twee methoden voor domain randomisation werden vergeleken: 1) het gebruik van synthetische afbeeldingen voor training met willekeurig contrast, gegenereerd uit segmentaties van MRI scans en 2) trainen op willekeurige, lineaire combinaties van twee MRI sequenties. De beste aanpak wat betreft beeldgelijkenis tussen CT en sCT werd gekozen voor vergelijking met de Baseline modellen. In een eindvergelijking werd naast de beeldgelijkenis ook de nauwkeurigheid van sCT-gebaseerde bestralingsplannen beoordeeld.

Resultaten: Het Baseline model getraind met T1w(Gd) en T2w afbeeldingen behaalde een betere beeldgelijkenis dan een model dat werd getraind met alleen T1w(Gd) afbeeldingen. De methode voor domain randomisation bestaand uit het toevoegen van afbeeldingen met willekeurig contrast resulteerde in een betere beeldgelijkenis voor de validatie dataset dan het model getraind met lineaire combinatie afbeeldingen en werd geïmplementeerd. Van de modellen die in de eindvergelijking werden opgenomen, behaalde het Baseline model de slechtste resultaten voor FLAIR, met een mean absolute error (MAE) van 106 $\pm$ 20,7 HU (gemiddelde $\pm\sigma$). De resultaten voor FLAIR behaald door het Domain Randomisation model waren significant beter met MAE = 99,0 $\pm$ 14,9 HU. Desalniettemin waren deze resultaten ondergeschikt aan die behaald door het Baseline+FLAIR model, getraind met toevoeging van FLAIR afbeeldingen aan de trainingsdata (MAE = 72,6 $\pm$ 10,1 HU). Zowel het Domain Randomisation model als het Baseline+FLAIR model resulteerde in een lichtelijk verhoogde MAE voor de geziene sequenties vergeleken met het Baseline model. De 3D γ-pass rates waren $>$ 95 % voor alle modellen en sequenties. Het 3D γ-pass rate met 1%,1mm criterium verkregen voor FLAIR afbeeldingen voor het Domain Randomisation model was significant hoger dan dat vergkregen voor het Baseline model (99,2 $\pm$ 0,9 % vs 99,0 $\pm$ 1,1 %), doch lager dan dat verkregen voor het Baseline+FLAIR model (99,4 $\pm$ 0,8 %). De pass rates voor de geziene sequenties verkregen voor het Domain Randomisation model verschilden niet significant van die verkregen voor het Baseline model.

Conclusies: Zelfs zonder domain randomisation werd een acceptabele dosimetrische nauwkeurigheid gevonden wanneer getraind werd op een mengsel van sequenties, zelfs voor een ongeziene sequentie. Niettemin resulteerde domain randomisation in verbeterde prestaties (beeldgelijkenis en dosimetrische nauwkeurigheid) voor de ongeziene sequentie, vergeleken met een model dat alleen op niet-synthetische MRI werd getraind. Dit resultaat wijst erop dat de methode zou kunnen helpen de noodzaak om netwerken te her-trainen te reduceren wanneer een model moet worden gebruikt voor een sequentie die geen deel uitmaakte van de trainingsdata.

Trefwoorden

Arteficiële intelligentie, Domain randomisation, Generaliseerbaarheid, GAN, Generative adversarial network, Image-to-image translatie, MRI-geleide radiotherapie, Synthetische CT

Contrast generalisation in deep learning-based brain MRI-to-CT synthesis

Lotte Nijskens, *M.Sc. student Biomedical Engineering, Imaging & In Vitro Diagnostics*

Abstract

Background and purpose: Computed tomography (CT) is the basis for radiotherapy (RT) planning, providing information on electron density needed for dose calculations. Magnetic resonance imaging (MRI) has superior soft tissue contrast and is helpful in tumour delineation, but, unlike CT, it does not inherently contain electron density information. Consequently, CT and MRI are often combined in the planning workflow, requiring image registration. MR-only based RT has been proposed to avoid residual errors after registration, reduce patients' exposure to ionising radiation and simplify the workflow. Synthetic CT (sCT) must be generated to enable MRI-based RT planning, which is possible within seconds using deep learning (DL). MRI acquisition protocols may change over time or differ between centres. Without network re-training, DL models poorly generalise to new domains, including different acquisition protocols, hindering their widespread implementation. Domain randomisation is a learning method that involves generating training data with randomised parameters. The method showed promising results for improving generalisation in, e.g., a segmentation task. This work investigates the ability of DL models for brain sCT synthesis to generalise to MRI scans acquired with unseen sequences without network re-training and how domain randomisation affects model performance on unseen sequences.

Materials and methods: Data from 95 patients undergoing RT were included from a retrospective database, requiring a CT image and corresponding T_1-weighted MRI with and without contrast (T1wGd and T1w), T_2-weighted (T2w) and FLAIR MRI. A Baseline conditional generative adversarial network was trained with and without an unseen sequence (FLAIR) to test how a model performs on the unseen sequence without domain randomisation. Also, two domain randomisation approaches were compared: 1) using synthetic training images with random contrast generated from segmentations of acquired MRI and 2) training on random linear combinations of two MRI sequences. The best approach regarding image similarity between sCT and CT was chosen for comparison with the Baseline models. In a final comparison, image similarity and accuracy of sCT-based dose plans were assessed.

Results: The Baseline model trained on T1w(Gd) and T2w images achieved better image similarity on the validation set's FLAIR images than a model trained only on T1w(Gd) images. The domain randomisation method of adding random contrast images resulted in better image similarity on the validation set than the model using linear combination images and was adopted. Of the models included in the final comparison, the Baseline model had the poorest performance on FLAIR, with mean absolute error (MAE) = 106 $\pm$ 20.7 HU (mean $\pm\sigma$). Performance on FLAIR significantly improved for the Domain Randomisation model with MAE = 99.0 $\pm$ 14.9 HU. Still, it was inferior to the performance of the Baseline+FLAIR model, trained by adding FLAIR images to the training set (MAE = 72.6 $\pm$ 10.1 HU). The Domain Randomisation and Baseline+FLAIR models resulted in a slight increase in MAE on the seen sequences compared to the Baseline model. The 3D γ-pass rates were $>$ 95 % for all models and sequences. The 3D γ-pass rate with 1%,1mm criterion obtained for the Domain Randomisation model for FLAIR images was significantly higher than that obtained for the Baseline model (99.2 $\pm$ 0.9 % vs 99.0 $\pm$ 1.1 %), yet lower than that obtained for the Baseline+FLAIR model (99.4 $\pm$ 0.8 %). Differences in pass rates obtained for the seen sequences between the Domain Randomisation and Baseline model were insignificant.

Conclusions: Even without domain randomisation, a satisfactory dosimetric accuracy could be obtained when training on a mix of acquired sequences, even for an unseen sequence. However, domain randomisation improved performance (image similarity and dose accuracy) on the unseen sequence compared to a model trained only on acquired MRI, indicating that the method could help reduce the need for network re-training if the model is to be used on a sequence unseen during network training.

Keywords

Artificial intelligence, Domain randomisation, Generalisation, Generative adversarial network, Image-to-image translation, MR-guided radiotherapy, Synthetic CT

I. Introduction

A. Radiotherapy planning

Radiation therapy (RT) is one of the main pillars of cancer treatment, indicated for approximately half of all cancer patients [1]. Computed tomography (CT) images are the basis for RT planning, as they provide the information on electron density required for dose calculations [2]. However, intra- and interobserver variability in delineating tumours and organs-at-risk (OARs) on CT are high, with the delineation of the gross tumour volume (GTV) having been called the 'weakest link' in RT accuracy [3].

Magnetic resonance imaging (MRI) has superior soft tissue contrast compared to CT. Hence, MRI has been suggested as the preferred imaging modality for delineating tumours and the surrounding OARs in RT planning [4]. Adding MRI to the treatment planning protocol can significantly reduce the intra- and interobserver variability in the delineation of tumours and OARs for multiple disease sites, including the breast [5], prostate and head and neck [6]. Additionally, for certain brain cancer patients, MRI can resolve tumour boundaries unidentified on CT [7, 8].

Unlike CT, MRI does not inherently contain the information needed for dosimetric computations [9]. On the one hand, in CT, the electron density in the imaged tissues determines the external radiation beam attenuation and image contrast. On the other hand, MRI is a resonance phenomenon occurring for atomic nuclei with an odd number of protons (or neutrons) that possess a non-zero nuclear spin, which is not directly related to electron density [10]. Consequently, CT and MRI are combined in the RT planning workflow whenever MRI provides additional information. The tumour and OARs are delineated on MRI, and this target definition is translated to CT through image registration [11], thereby introducing an additional uncertainty factor into the workflow [12, 13]. Differences in patient positioning or bladder or rectal filling, caused by the image acquisition at different time points, complicate the registration process [14].

B. MR-only treatment planning

MR-only based RT has been proposed to avoid the error associated with image registration [15, 16]. Removing the registration step reduced systematic uncertainties for the prostate from 3-4 mm in

a CT/MRI hybrid pathway to 2-3 mm in an MR-only planning method [14]. Also, MR-only RT reduces the patient's exposure to ionising radiation [11], which can be of significant benefit when re-planning is needed [17] or for paediatric populations [18]. Another advantage is that fewer scans are required, improving patient comfort [19]. Moreover, eliminating the CT scan and simplifying the workflow reduces the workload [20, 21] and costs [19, 20, 22].

With the clinical introduction of MRI scanners integrated with linear accelerators for radiation delivery for MR-guided radiotherapy (MRgRT) [23, 24], MR-only RT became particularly interesting. Because in MRgRT, MRI can be acquired before, during, and after the RT delivery, both the planning and treatment phases are MRI-guided [25]. MRgRT even allows online re-planning or plan adaptation if changes in patient anatomy are observed [24, 26].

C. Synthetic CT (sCT) generation

The lack of a physical relationship between the tissues' nuclear magnetic properties and their electron density characteristics needed for dose calculations can be regarded as the main hurdle in implementing MR-only RT. Many approaches for representing MRI as a CT equivalent have been proposed to overcome this problem [27]. This process is called synthetic CT (sCT) generation. Alternative names for the resulting image are pseudo-CT, MRCT, virtual CT, or substitute CT [20, 28].

Approaches for sCT generation can be divided into three main categories: atlas-based, voxel-based and hybrid [19]. In atlas-based approaches, the voxels within a patient's MRI scan are (deformably) registered to a CT atlas or an MRI atlas for which there is a known correlation with a given quantity, mainly Hounsfield units (HU) [19]. The accuracy of atlas-based methods depends on the registration accuracy. Multi-atlas approaches can be beneficial in this respect, offering a broader range of options for registration than approaches using a single (average) atlas [20]. Generally, atlas-based methods do not perform well in the case of atypical anatomies or less common patient populations [29, 30].

Voxel-based approaches use the intensity and structure of an MRI for conversion to electron density [19]. Such methods often rely on specialised MRI sequences to separate bone and air, like ultra-short echo time sequences [30]. Advantages of voxel-based methods include the lack of need for image registration and segmentation when considering statistical approaches [30]. Compared to atlas-based approaches, voxel-based techniques handle atypical anatomies better [30, 31]. An important subcategory of voxel-based methods are methods based on machine learning or, more recently, deep learning (DL) [19]. In image synthesis tasks (e.g., sCT generation), a DL model translates or maps a source to a target image belonging to a different imaging domain [32]. Unlike other voxel-based methods [30], learning-based methods can separate bone and air without specialised MRI sequences [31].

Hybrid approaches combine elements from voxel- and atlas-based techniques [30]. Sometimes, although also viewed as a subcategory of voxel-based methods [19], a fourth category is distinguished: bulk-assignment approaches. A (coarse) segmentation is performed on the MRI, and each label is assigned a value in HU [28, 30]. Its simplicity makes this an appealing method [30]. Clinically, however, this method is considered the least advantageous compared to atlas-based and voxel-based approaches [30]. Treatment planning can lead to dose differences above 2% with the acquired CT image [30], the suggested limit for clinical acceptability [33].

Regarding accuracy, atlas-based and learning-based methods show promise and are actively researched [34]. DL-based methods are of particular interest as they require a limited time for sCT generation once training is finished, unlike atlas-based methods [35]. Whereas (multi-)atlas-based methods can be computationally expensive, with the time needed for sCT generation ranging from ten minutes to multiple hours, DL models can generate an sCT in seconds to a minute [34, 35]. This time aspect is essential in MRgRT, requiring sCT generation within minutes to allow daily re-planning [26, 28].

D. Deep learning-based sCT

Convolutional neural networks (CNNs) are the most common and most successful type of DL model in (medical) image processing [36], including image synthesis tasks [27]. CNNs consist of layers of convolutional filters combined through weights and biases. Model training involves optimising these parameters, a process guided by a loss function that measures the error between the network output and ground truth for a given set of parameters. The network calculates a gradient vector representing, per weight, how much this error changes for a slight increase/decrease (determined by the learning rate) of that weight. Model parameters are then updated based on this gradient vector to minimise the loss function [37].

In 2016, Nie *et al.* [38] proposed using a 3D fully convolutional neural network for the generation of sCT from MRI for the first time. After the introduction of generative adversarial networks (GAN) [39], Nie *et al.* [40] implemented for the first time a conditional GAN (cGAN). Subsequently, Wolterink *et al.* [41] proposed in 2017 an unsupervised CycleGAN [42] for sCT generation. Since then, DL networks for sCT generation have been trained and tested for a multitude of anatomical sites, including the abdomen [43, 44], brain [41, 45], breast [46, 47], head and neck [34, 45], liver [44], pelvis [48, 49], prostate [43, 50] and rectum [51].

E. Generalisation of DL models for sCT generation

DL models are known to poorly generalise to new domains [52-54]. Models assume a shared statistical distribution and feature space between the training and test data. Therefore, they must be re-trained if the distribution of the test data changes [52]. For sCT, new domains could be, e.g., a different MRI sequence, an MRI acquired in another hospital with different acquisition parameters, MRI acquired after a scanner upgrade or with a different model, or another anatomy.

Most DL models for sCT generation were trained and tested on a specific anatomical site using MRI acquired with a limited number of sequences and a fixed range of imaging parameters. These models do not consider the variability in MRI acquisition protocols employed in clinical practice or protocol changes that might occur over time. If imaging parameters are changed compared to the training data, networks may need to be re-trained on newly acquired data, hindering their use in clinical practice. Models that can generalise to MRI sequences unseen during training, would avoid re-training and facilitate a widespread clinical implementation [49]. Suppose generalisation to a different MRI sequence is possible without network re-training. The model can likely bridge a smaller contrast domain gap, e.g., renewed acquisition parameters, MRI scanners from different vendors, or data from other hospitals [55]. Additionally, models would no longer require a fixed input sequence, which is helpful if non-standard imaging protocols are acquired for specific patients, e.g., because of claustrophobia or contrast allergies.

Two recent publications investigated generalisation to multiple MRI sequences in MR-only RT [56, 57]. The methods employed in these works to improve generalisation involved (re)training the network on those sequences. Zimmermann *et al.* [57] showed it is possible to train a combined model for T1w(Gd) and T2w images. A comparison with single-sequence models revealed poor generalisation to the sequence not included in the training data [57]. Similar poor generalisation to an unseen sequence was found in [56].

F. Domain randomisation

Recently, a promising technique was proposed to improve a segmentation network's ability to generalise to unseen MRI sequences [58, 59]. Motivated by evidence that data augmentation beyond realism improves generalisation [60], domain randomisation is a learning method that involves generating training data with randomised parameters (e.g., colour). The method relies on the hypothesis that elaborate variability in synthetic training data causes the model to see reality as a variation of the training data [60] and tries to bridge all domain gaps, in this case in MRI space, at once [55].

Billot *et al.* [59] applied domain randomisation to train a CNN to automatically segment brain images of any type of MRI contrast and any resolution. Instead of using images for each type of contrast during training, Billot *et al.* [58, 59] used segmentations to create synthetic images with a broad, not per se realistic, range of contrasts (Fig. 1). The segmentation CNN was trained solely on synthetic images to remove the bias for certain types of contrast, claiming that the model learns contrast-agnostic features [59]. The results have been promising, and the method has successfully been extended to other tasks by the same group [61, 62], but not sCT generation. Given the promising results on the other tasks, it would be interesting to see if the method may also be applied to the sCT generation task.

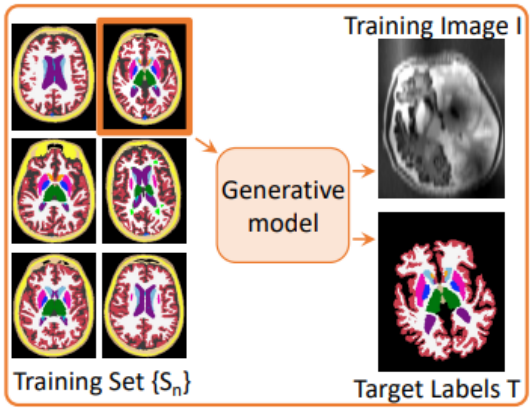

Fig. 1. Example of the synthetic training data used in [59]. A segmentation is converted to a synthetic image by assigning a Gaussian distribution with a random mean and standard deviation to each label and filling the voxels within the label according to this distribution. Additional augmentation steps include elastic deformation. From: [59].

G. Research goals

This work investigates the ability of DL models for MRI-to-sCT generation to generalise to MRI scans acquired with unseen sequences in MR-only RT. We concentrate on one anatomical region: the brain. Encouraged by the results obtained in [48] and [63], we focus on the cGAN framework termed `pix2pix` [32].

Inspired by [58, 59], we propose a domain randomisation method that converts the MRI in the training dataset into synthetic images of random contrast to improve contrast generalisation. We hypothesise that training a DL model for MRI-to-sCT generation on input data with synthetic, not necessarily realistic, image contrast obliges the network to learn contrast-agnostic features. We formulated the following research question to test our hypothesis: How does a domain randomisation strategy of randomising image contrast in the training dataset influence the performance, in terms of image similarity and dosimetric accuracy, on unseen MRI sequences of a cGAN model tasked with sCT generation from brain MRI?

Two approaches to domain randomisation are investigated: 1) using synthetic training images created from label maps of the acquired MRI; 2) training on random linear combinations of two MRI sequences. The effect of mixing acquired sequences in the training data on generalisation is also studied. The end goal is to investigate how far a DL model can be pushed towards becoming contrast-agnostic, capable of generating sCT from which RT plans can be calculated with clinically acceptable dosimetric accuracy.

II. Related work

Several previous studies have investigated the generalisation ability of DL models for sCT generation or a related task. The following categories were identified and are discussed in the following sections: dataset balancing, transfer learning, data augmentation, domain randomisation and other approaches. We summarised the categories' (dis)advantages in the context of generalisation to unseen sequences (Table I).

A. Dataset balancing

Several publications investigated the impact of introducing variation in the training data. We have grouped those studies investigating model performance on different types of input data under the term dataset balancing. Bird *et al.* [51] found no significant differences in cGAN performance on anorectal T_2-weighted MRI data from two different centres. Similarly, in [64], a cGAN generated clinically acceptable sCT images from T_2-weighted pelvic MRI, generalisable to MRI scanners from multiple vendors and data from a held-out centre. In [65], model robustness for independent data from an external centre was shown for a 3D U-Net trained for sCT generation for a head-and-neck cohort. Fetty *et al.* [66] demonstrated cGANs' ability to generalise to MRI scanners of different field strength without finetuning models on data from these other scanners. Maspero *et al.* [63] studied the influence of heterogeneity in a patient population and imaging protocol on sCT image quality from T_1-weighted paediatric brain MRI. Image quality was independent of scan parameters, field strength, Gadolinium administration, patient age, size and shape [63]. This group also demonstrated the feasibility of employing a single cGAN for sCT generation for the entire pelvis, despite training solely on prostate images [48]. The feasibility of using multiple MRI sequences for training a combined model for brain sCT generation was investigated in [57], showing that a combined model for T_1- (with and without contrast) and T_2-weighted images can achieve performance comparable to sequence-specific models for each sequence.

Altogether, introducing variability in training data improves generalisation effectively when target data are added to the training set. Additionally, relatively small domain gaps, e.g., same-sequence data from other centres, can be bridged without re-training. Without re-training, performance generally decreases for larger domain gaps, like unseen sequences [56, 57].

B. Transfer learning

Transfer learning is a domain adaptation method often used to overcome the problem of small available training datasets: an existing model that performs well after training on a large dataset is adjusted to a new target domain. Thereby, the method generally requires less target domain data than dataset balancing. For instance, Wang *et al.* [67] applied transfer learning to finetune a model for MRI-to-sCT generation from the paediatric brain to the paediatric pelvic region, for which only limited data were available. They found improved performance on the body surface and in bone compared to training a pelvic model from scratch. In [65], transfer learning was successfully applied to adapt a 3D U-Net for sCT generation to new data acquired

after a software update. Li *et al.* [56] explicitly investigated how training strategy influenced networks' ability to generalise to brain data from other centres and MRI sequences. The most beneficial approach consisted of pre-training on source data and re-training with target data of a different sequence. A model trained only on source data resulted in the poorest performance on the target data, with results being unsuitable for dose calculations [56].

C. Data augmentation

Data augmentation methods increase the variability in the dataset available for training. The method is designed to improve the generalisability of DL models and can aid in bridging (small) domain gaps [68]. Basic data augmentation methods, such as intensity augmentations, random flipping, and randomly cropping the input images, have been applied in multiple studies on sCT generation from MRI (e.g., [41, 48]). However, these studies did not explicitly investigate the effect of such methods. In [69], a more elaborate data augmentation method was applied: MRI from two patients were combined with Laplacian blending to simulate a larger training cohort, improving image similarity between sCT and acquired CT on the test set [69].

D. Domain randomisation

Domain randomisation, introduced in I-F, seeks to bridge larger domain gaps than data augmentation. As mentioned in section I-F, Billot *et al.* [59] applied domain randomisation to achieve contrast-agnostic segmentation of brain MRI. The same group also applied this learning strategy to an image registration task [61] and for generating super-resolution MRI from lower resolution images [62].

E. Other approaches

Two other approaches were identified to improve model generalisation, specifically for supervised frameworks. The CycleGAN model for unsupervised training learns mappings by enforcing cycle consistency, meaning the mapping function from one domain (e.g., MRI) to the other domain (e.g., CT) should be reversible [42]. This cycle consistency poses inherent problems with generalisation for the CycleGAN architecture because it assumes a one-to-one relationship between the two domains [70]. Brou Boni *et al.* [49] addressed this issue by adapting the Augmented CycleGAN (AugCGAN) model proposed by Almahairi *et al.* [70]. The model synthesised sCT with clinically acceptable dosimetric accuracy from MRI acquired with different scanners and scan parameters and generalised to data from a centre excluded from the training set [49]. Alternatively, Gadermayr *et al.* [71] introduced an asymmetric cycle consistency loss in the original CycleGAN architecture to create pseudo-healthy MRI from thigh MRI with fat-infiltration [71].

III. Materials and Methods

Networks for sCT generation were trained using either 2D sagittal slices or isotropic 3D patches. After model optimisation, a 2D or 3D configuration was chosen for all subsequent experiments.

A Baseline model with and without an unseen sequence was trained to test how a model generalises to the unseen sequence without domain randomisation. Also, generalisation was investigated by comparing two domain randomisation approaches: 1) using synthetic training images with random contrast; 2) using randomly generated linear combinations of acquired images as training data.

This section first describes the general methodology: data collection and acquisition, image processing and performance evaluation. Then, the specifics of the network architecture, model optimisation, the two methods for domain randomisation and the experimental set-up are explained, and details on the final comparison are provided.

TABLE I
Advantages and disadvantages of the different approaches to improve generalisation in sCT generation.

Category	Advantages	Disadvantages
Dataset balancing	• Simple implementation • Effective	• Requires including target data in the training set[a] • Poor results for target data outside the training set for large domain gaps
Transfer learning	• Simple implementation • Effective • Requires smaller amounts of target data than dataset balancing	• Requires including target data in the training set[a] • Poor results for target data outside the training set for large domain gaps
Data augmentation	• Increases variability in available training data • Effective for bridging small domain gaps	• Effectiveness not expected for large domain gaps
Domain randomisation	• Aims to generalise to data outside the target domain • Aims to bridge larger domain gaps	• Requires generating synthetic training data
Other approaches		• Described methods are only applicable for unsupervised training

[a]Unsuitable when investigating generalisation to an unseen sequence.

A. Data collection and acquisition

This study was conducted under the local Medical Ethical Committee (study number: 20/519, approved on August 11, 2020). Data were selected from a retrospective, anonymised database of patients undergoing treatment at the UMC Utrecht RT department. The main selection criterion was the availability of a treatment plan for brain RT conducted between January 2020 and July 2021, with corresponding CT and MRI (T_1-weighted with and without contrast enhancement, T_2-weighted and FLAIR images). Patients were excluded if not all sequences were available, no suitable CT was available, the time between MRI and CT acquisition was more than 1.5 months, patient age was $<$ 18 years, or the MRI was a follow-up exam. If multiple CT acquisitions were available, the most recent one was chosen, and the MRI dataset acquired closest in time to the CT was selected.

In total, 95 patients were selected. These were randomly divided over the training (n = 60), validation (n = 10) and test set (n = 25). The female/male ratio for the 95 included patients was 51/44. The mean patient age was 59.9 ±13.0 years (range: 24.3-86.8). The mean interval between CT and MRI acquisition was six days (1-26). Dose prescriptions ranged from 14 to 60 Gy over 1-33 fractions.

Planning CTs were acquired at the radiotherapy department using a Brilliance Big Bore system (Philips Healthcare, USA). The acquisition took place in the supine treatment position, aided by head support and a personalised immobilisation mask. CT acquisition was without contrast agents, with a tube potential of 120 kV, a tube current of range 234-360 mA, and 1000-1712 ms exposure. The in-plane resolution was 0.57-1.17 mm^2, with a slice thickness of 1-2 mm.

MRI data were acquired with a 1.5 or 3.0 T Ingenia MR-RT system (Philips Healthcare, the Netherlands). Available sequences were: 3D T_1-weighted turbo field echo (TFE) images with and without Gadolinium contrast (T1w and T1wGd), 2D T_2-weighted turbo spin-echo (TSE) images with Gadolinium contrast (T2w) and 3D T_2-weighted FLAIR TSE images (FLAIR). Table II gives an overview of the acquisition parameters per sequence.

TABLE II
OVERVIEW OF ACQUISITION PARAMETERS PER SEQUENCE FOR THE 95 INCLUDED PATIENTS.

Parameter	Value(s)			
	3D T1w TFE	**3D T1w TFE with Gadolinium**	**2D T2w TSE with Gadolinium**	**3D T2w FLAIR TSE**
Field strength B_0 [T]	**1.5**: n = 66 **3.0**: n = 29	**1.5**: n = 66 **3.0**: n = 29	**1.5**: n = 66 **3.0**: n = 29	**1.5**: n = 66 **3.0**: n = 29
Contrast	No	Yes, Gadolinium	Yes, Gadolinium	No
Read-out direction	Anterior-posterior	Anterior-posterior	Anterior-posterior	Anterior-posterior
Flip angle [°]	8	8	90	90
Repetition time (TR) (range) [ms]	7.6-8.7	7.6-8.7	3119-5996	4800
Echo time (TE) (range) [ms]	3.5-4.1	3.5-4.1	80-100	303-363
FOV[a] (range) [mm^3]	230, 230, 160	230, 230, 160	230, 230, 140-160	230, 230, 160
Acquired voxel size[a] (range) [mm^3]	1.0, 1.0, 0.5-1.0	1.0, 1.0, 0.5-1.0	0.6, 0.6-0.7, 4.0-5.0	1.1-1.2, 1.1-1.2, 0.6
Reconstructed voxel size[a] (range) [mm^3]	0.4-1.0, 0.4-1.0, 0.5-1.0	0.5-1.0, 0.5-1.0, 0.5-1.0	0.4-0.5, 0.4-0.5, 4.0-5.0	1.0, 1.0, 0.6
Reconstruction matrix[a] (range)	240-512, 240-512, 162-323	240-480, 240-480, 162-323	480-512, 480-512, 31-43	240, 240, 269-270
Bandwidth (range) [Hz/px]	190-217	190-217	143-206	851-1075
Acquisition time (range) [s]	136-271	121-271	117-137	331-475

[a]Respective directions: anterior-posterior, right-left and cranio-caudal.

B. Image processing

1) Pre-processing: Each MRI was rigidly registered to the corresponding CT with `Elastix` software (version 4.700) [72, 73], using multi-resolution registration (with resolution σ = 4, 2, 1 and 0.5) with an adaptive stochastic gradient descent (ASGD) optimiser and mutual information similarity metric. The parameters from [63] were adopted. The registered MRI will be referred to as MRI_{reg}. MRI_{reg} and CT were resampled to isotropic 1.0x1.0x1.0 mm^3 resolution using linear interpolation.

Matlab R2019a (The MathWorks, Inc., USA) was used for the subsequent pre-processing steps. Most images contained a discrepancy between CT and MRI FOVs, caused by angulated MRI acquisition (Fig. 2). A binary body mask was computed on the non-registered MRI to ensure congruent FOVs between CT and MRI_{reg}. The mask was generated using a threshold with an empirically determined value of 20 (or 15 for T2w images), followed by morphological filling and dilation with a disk-shaped structuring element of radius 20 voxels. The binary mask was registered to the CT by applying the transform computed for MRI_{reg}, resampled, and applied to the MRI_{reg}-CT pair for training. The FOVs of MRI_{reg} and CT were cropped to the extent of the registered mask with additional ten voxel margins on each side or until the original image boundary. MRI_{reg} were normalised by clipping to the per-patient 99th percentile value over the masked volume. Training CTs were clipped to range [-1024, 1500] HU.

Note that for CT, the masking and range clipping steps were only applied to the training images (hereafter: CT_{train}). Fig. 2 shows an example of a normalised brain MRI_{reg} with the corresponding normalised CT_{train}, ground truth CT_{crop} and original, unregistered MRI for each sequence for a single patient. CT_{train} and normalised MRI_{reg} were saved as 2D sagittal slices in PNG format, linearly rescaled to 16-bit unsigned integers, and as 3D volumes in NifTI format, linearly rescaled to [-1, 1].

For the experiment investigating how training on linear combination images affects model generalisability, the pre-processed T2w and T1wGd images were also registered to the corresponding T1w image with the same registration parameters as before.

2) Post-processing: Image post-processing was done in Matlab 2019a. Inference of models with 2D configuration generated 2D sagittal sCT slices. These were linearly resampled and stacked to obtain volumes of 1.0x1.0x1.0 mm^3 resolution. Inference of models with 3D configuration generated 3D volumes of the required resolution without additional steps. All generated sCTs were linearly rescaled to a [-1024, 1500] HU range, conforming to the range of CT_{train}.

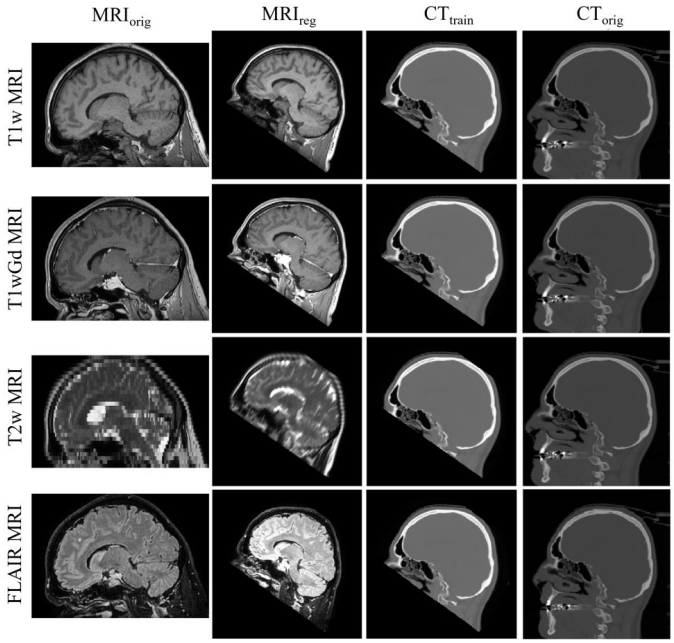

Fig. 2. Example of the pre-processing outcomes. The original T1w (top row), T1wGd (second row), T2w (third row) and FLAIR (bottom row) brain MRI for a single male patient in the training dataset are shown (left) with corresponding normalised MRI_{reg}, CT_{train} and ground truth CT_{crop} (left to right).

C. Performance evaluation

Matlab 2019a was used for performance evaluation. The experiments evaluated image similarity between acquired CT and generated sCT for the validation set. Image similarity metrics and dosimetric accuracy were calculated on the test set for models included in the final comparison. Statistical comparisons were performed with

Wilcoxon signed-rank tests, with p-values < 0.05 regarded as statistically significant.

1) Image similarity: The accuracy of the assigned HU values was analysed with a voxel-wise comparison between CT_{crop} (ground truth) and sCT. A body contour mask was applied to CT_{crop} and sCT before calculating the metrics, comparing over the intersection of the two masks. The masks were created by thresholding the (s)CT above -200 HU, then morphologically closing and filling the combined mask to include the nasal cavities. The mean absolute error (MAE) was computed per patient. Peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM) were additional metrics. Definitions are provided in Appendix A. The range and mean $\pm$ standard deviation ($\mu \pm \sigma$) over all patients in the validation or test set were calculated for each metric.

2) Dosimetric accuracy: For patients in the test set (n = 25), the clinically optimised dose plan was re-calculated on (s)CT. Generated sCTs were registered and resampled to the original, non-cropped CT with `Elastix` (version 4.700) [72, 73], allowing only translations. Multi-resolution registration was performed (with isotropic resolution σ = 4, 2, 1 and 0.5) with an ASGD optimiser and mutual information. In some cases, three resolutions were used instead of four (resolution 1: σ = 4, 4, 2; resolution 2: σ = 2, 2, 1; resolution 3: σ = 1, 1, 0.5).

A segmentation of the body contour of the non-cropped CT was taken from the clinical treatment plan, and the voxels outside the original MRI FOV and inside this body contour were set to 0 HU. The difference in FOV between sCT and acquired CT was thus water-filled in both images (example in Appendix B). The water-filled sCT and acquired CT are referred to as sCT_{wf} and CT_{wf}.

Plan re-calculation was done on (s)CT_{wf} using GPUMCD [74]. Plans were volumetric modulated arc therapy (VMAT) photon plans with a single arc with a beam energy of 6.0 MV. They were calculated with a Monte Carlo algorithm on a 3 mm^3 grid with 3 % uncertainty.

Dosimetric performance was assessed through the calculation of the dose difference (DD) relative to the prescribed dose (D_{presc}) in the high dose region (D $>$ 90% of D_{presc}) [27]:

$$DD = 100 * \frac{D_{CT} - D_{sCT}}{D_{presc}}\%, \quad (1)$$

with D the dose (in Gy) in the (s)CT_{wf}-based dose plan. Korsholm *et al.* [33] proposed a criterion for clinical acceptability of DD: the DD compared to a CT-based dose plan should be $<$2% for 95% of the patients. In this work, a more conservative criterion was adopted. Individual sCTs were considered acceptable if the DD was $<$1%.

Dose-volume histograms (DVH) were analysed for differences in D_{median} and D_{max} between sCT- and CT-based plans for the following OARs: brainstem, optic chiasm, lenses, cochleae and pituitary gland. Additionally, a 3D-γ global analysis was conducted [75]. The definition of the γ-index and -pass rate is given in Appendix C. For the computation of γ-pass rates, a 10% dose threshold was used, with 3%,3mm, 2%,2mm and 1%,1mm criteria. Heilemann *et al.* [76] demonstrated the ability to detect clinically unacceptable VMAT plans using a 90% γ-pass rate threshold for the 2%,2mm criterion. Nevertheless, the absence of clinically significant dose differences was not guaranteed [76]. Considering also that evaluation of γ-pass rates is adopted for quality assurance of delivered plans, where uncertainty is higher, we adopted stricter thresholds in this work: 95% and 99% for the 2%,2mm and 3%,3mm criteria.

D. Network architecture

The cGAN model `pix2pix` was implemented to allow paired training, as proposed by Isola *et al.* [32]. For models with a 2D configuration, a PyTorch version 1.4 implementation of the original `pix2pix` model [32] was used, modified to enable training on 16-bit greyscale images. An implementation of the original `pix2pix` model [32] called `Ganslate` [77] in PyTorch version 1.10 was used for 3D models. All models were trained using a GPU (Tesla P100 PCIe 16 GB or V100 PCIe 32 GB, NVIDIA Corp., USA).

Figure 3 illustrates the general structure of the implemented cGAN. The model's goal is to learn a mapping from an image (x) belonging to a specific input domain (MRI) to an output image (y') from a different domain (y; here: CT). The model consists of two networks: a generator (G) and a discriminator (D). G aims to produce realistic images (y'; the sCT) resembling the example images from the target domain (y). D is presented with images from the target domain (y) and images output by G (y') and tries to identify y'. The two networks are trained in an adversarial manner and compete in a minimax game. While G learns to produce better images, D becomes better at discriminating between real and fake [32].

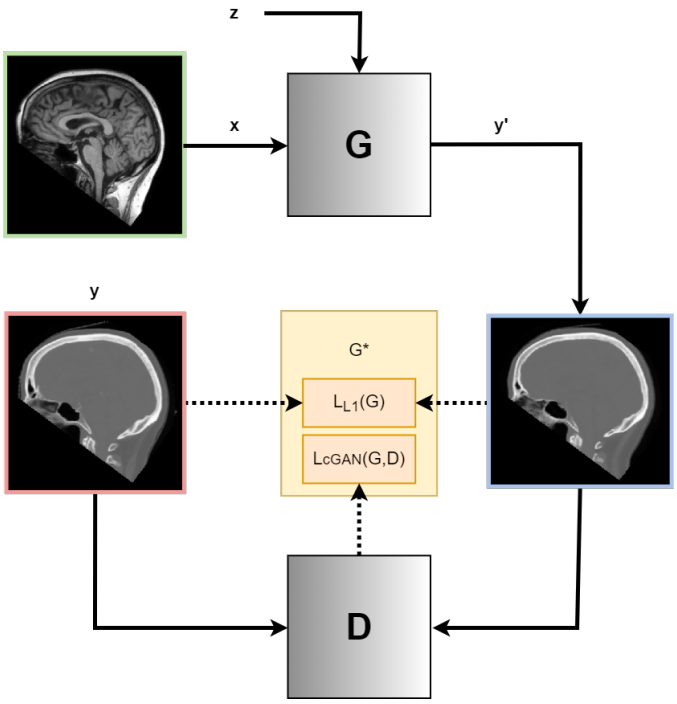

Fig. 3. Schematic overview of the employed cGAN for sCT generation. The model consists of a generator (G) and a discriminator (D) competing in a minimax game. G tries to produce images y' (sCT) from x (MRI) that are indistinguishable from their real counterpart (y; CT), while D is presented with y' and y and tries to discriminate between real and fake. Random noise (z) is injected in the form of dropout layers in G. G* is the objective of the learning process consisting of an adversarial part (L_{cGAN}(G, D)) and a voxel-wise part (L_{L1}(G)).

Models with a 2D configuration were implemented with a 256 x 256 U-Net generator [32]. Models with a 3D configuration were implemented with a 3D U-Net generator architecture that allows variable patch sizes as input. A 70 x 70 PatchGAN discriminator [32] was used for both configurations. The L1-based loss function proposed in [32] was implemented as the following objective function (eq. 2).

$$G^* = arg \min_G \max_D L_{cGAN}(G, D) + \lambda L_{L1}(G) \quad (2)$$

This final objective consists of an adversarial part (L_{cGAN}, eq. 3) dependent on both G and D and a voxel- or pixel-wise term (L_{L1}, eq. 4) that depends only on G. The factor λ weighs the two terms. Isola *et al.* [32] have shown the importance of this combined loss

function. Using only L_{cGAN} produced sharp images at the cost of introducing artefacts. Solely using the L_{L1} term, on the other hand, led to good mappings, but the generated images were blurry [32].

$$L_{cGAN}(G,D) = \mathop{\mathbb{E}}_{x,y}[\log D(x,y)] + \mathop{\mathbb{E}}_{x,z}[\log(1-D(x,G(x,z))] \quad (3)$$

$$L_{L1}(G) = \mathop{\mathbb{E}}_{x,y,z}[\| y - G(x,z) \|]_1 \quad (4)$$

E. Model optimisation

A 2D and a 3D model were optimised in parallel to establish a model configuration for all subsequent experiments. Hyperparameter optimisation was performed with a grid search strategy using the validation set and ten patients from the training set, using only T1w images without contrast; see Appendix D for the details. The MAE served as the decision criterion. The optimised 2D and 3D model were trained on a subset of thirty patients from the training set. The performance of the two optimised models was compared (Appendix E), after which the 3D configuration was adopted. As a final optimisation step, the ratio between T1w images with/without contrast and T2w images in the training set was balanced, and the batch size was finetuned (Appendix F)[1]. The training dataset (n = 60) contained 60 T2w, 30 T1w and 30 T1wGd images after balancing.

After optimisation, all models were trained with Xavier initialisation, Adam optimiser, patch size = 128x128x128 voxels, batch size = 1, λ = 5000, number of downsampling steps = 5, and a constant learning rate of 0.001. A sliding window inferrer was used for patch combination with a patch overlap of 0.5 and Gaussian blend mode. The Adam optimiser [78] was implemented with β_1 = 0.5 and β_2 = 0.999 as momentum parameters and no weight decay. Early stopping was applied to avoid overfitting, as illustrated in Appendix G, using the MAE as the decision criterion.

F. Generating random contrast images

1) Segmentation: Automatic segmentations were generated to investigate how synthetic training images with random contrast created from label maps affect models' ability to generalise to unseen sequences. Segmentations were generated from the T1w images, complemented by some structures labelled using CT_{train}. Note: this research does not aim to train a segmentation network, and label maps do not serve as ground truth for comparisons.

Segmentation of cerebral structures was performed on T1w images using an open-source DL network called FastSurfer [79]. OARs were added by segmentation of T1w MRI with a previously in-house developed DL algorithm (unpublished and developed for clinical use as in [80]), based on the DeepMedic model [81]. The GTV was obtained from a clinical segmentation. Cerebrospinal fluid (CSF) was labelled using a combination of FastSurfer labels and a clinical segmentation. Thresholding CT_{train} segmented labels for bone and soft tissue. The resulting label maps (Fig. 4) were saved as an additional dataset. More details on image segmentation and a lookup table with included labels are reported in Appendix H.

2) Contrast randomisation: Label maps were converted to random contrast (RC) images for network training on the fly based on [59]. All steps involving generating RC images from label maps were implemented using TorchIO software [82].

After randomly selecting a segmentation from the training data, each label i in the label map was assigned a Gaussian function with mean and standard deviation chosen randomly from a uniform distribution with ranges of [10, 240] and [1, 25], respectively. These ranges were based on the sensitivity analysis conducted in [59]. All voxels within a label were assigned an intensity value sampled from this Gaussian distribution.

Then, images were blurred to increase spatial coherence between neighbouring voxels. The standard deviation of the Gaussian was randomly sampled from a uniform distribution: $\sigma_{blur} \sim U(0, 0.3)$, like in [58]. Random gamma augmentation was applied after rescaling image intensity to a positive range to increase variability in the training data further. Following [59], the exponent $\gamma = e^{\beta}$ was randomly sampled from a normal distribution: $\beta \sim N(\mu_\beta, \sigma_\beta)$ with $\mu_\beta = 0$ and $\sigma_\beta = 0.4$. The resulting RC image was rescaled to [-1, 1]. Figure 5 shows some example slices from patches of RC images used for network training, illustrating the variability in the training data generated with this method.

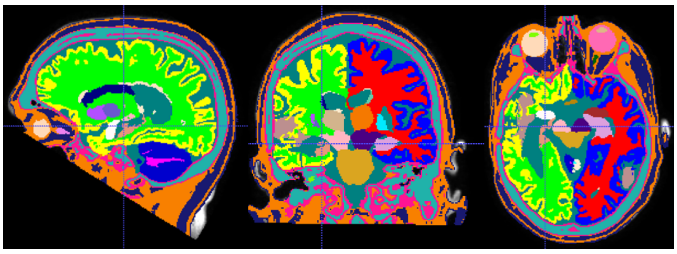

Fig. 4. Example of a label map created through automatic segmentation of a single patient's T1w MRI and corresponding CT.

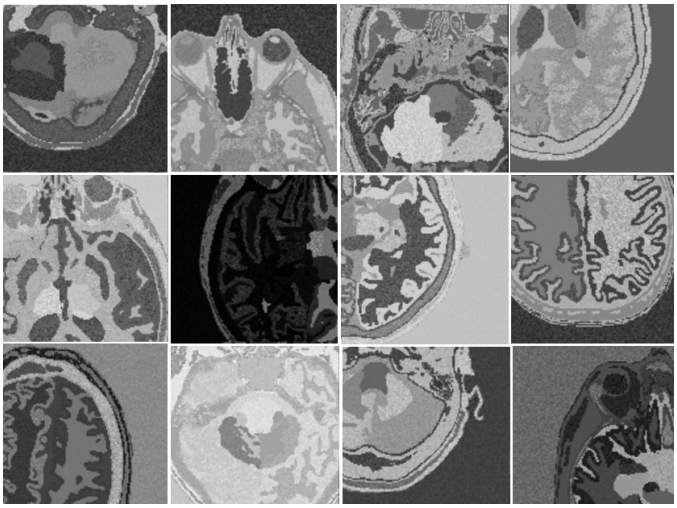

Fig. 5. Examples of random contrast (RC) images generated from label maps. Each image is a slice from an example patch as input to the network during training.

G. Generating linear combination images

The domain randomisation strategy comprising RC images (section III-F) requires segmenting patients' MRI. A second, more straightforward domain randomisation approach based on linearly combining acquired MRI was designed to prevent this need for label maps and test whether such a method could be effective.

Linear combination (LC) images were generated from T1w(Gd) and T2w images. A random choice was made during network training between combining the patient's T2w image with their T1w or T1wGd image, using equal probabilities. The T1w(Gd) and T2w

[1]One patient was retrospectively excluded from the validation set after failure of registration between the T2w MRI and the CT was observed. Validation of all models except the optimised 2D and 3D models was thus done on a nine-patient validation set.

images were then combined as follows:

$$Im_{LC} = p_1 * Im_{T1} + p_2 * Im_{T2}, \quad (5)$$

with p_1 and p_2 the coefficients for voxel-wise addition of the T1w(Gd) (Im_{T1}) and T2w (Im_{T2}) image, respectively. These were randomly sampled from a uniform distribution: $p_{1,2} \sim U(-1, 1)$. The chosen range allows addition and subtraction in the linear combination and contrast inversions. Finally, images were rescaled to the range [-1, 1]. Figure 6 shows several example LC image patches.

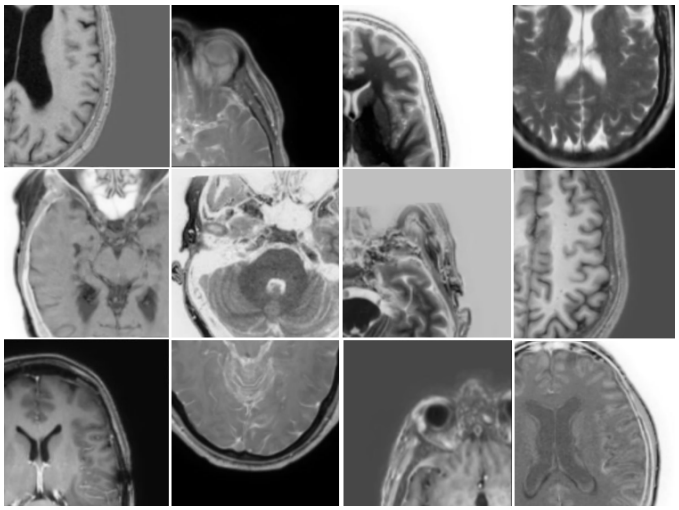

Fig. 6. Examples of linear combination (LC) images generated through a linear combination of real T1w(Gd) and T2w images. Each image is a slice from an example patch as input to the network during training.

H. Experiments

1) Generalisation of conventional cGANs - Baseline vs T1-only model: A Baseline model was trained on a mix of T1w, T1wGd and T2w images to assess models' ability to generalise to an unseen sequence (FLAIR) without data augmentation or domain randomisation. Its performance was compared to that obtained for a T1-only model to study the effect of mixing sequences on generalisation.

Table III indicates for which sequences the MAE was considered to decide when to apply early stopping and the chosen iteration, for each model trained for the experiments. The Baseline model was obtained by re-training the 3D model on the whole training set (n = 60) using 30 T1w, 30 T1wGd and 60 T2w images. The T1-only model was trained on the same dataset, excluding the T2w images. After early stopping, image similarity metrics were calculated for both models on sCT from T1w, T1wGd, T2w and FLAIR for the chosen iteration. The models' performances were statistically compared per sequence.

2) Domain randomisation - Random contrast vs mixing sequences: The RC+T1(Gd) model was trained to investigate the effect of adding RC images in the training data and compare it to the impact of mixing acquired sequences in the training data. The training dataset consisted of T1w images (n = 30), T1wGd images (n = 30) and segmentations (n = 60), which were converted on the fly to RC images as described in section III-F. After the choice of iteration by early stopping, image similarity metrics were additionally calculated for sCT generated from T2w and FLAIR images of the validation set. Results for the RC+T1(Gd) model were statistically compared with those obtained for the Baseline and T1-only model.

3) Domain randomisation - Random contrast vs linear combinations: This experiment studies whether using RC or LC images as training data affects models' ability to generalise to the unseen sequence (FLAIR) and results in choosing either of the two methods for the final comparison.

After comparison to a model trained on RC images only (Appendix I), the RC+T1(Gd)+T2 model was adopted for this experiment. Its training dataset consisted of a mix of label maps (n = 60) and acquired T1w (n = 30), T1wGd (n = 30) and T2w (n = 60) images. For the domain randomisation method comprising LC images, the LC+T1(Gd)+T2 model was adopted after comparison to the LC-only model (Appendix J). This LC+T1(Gd)+T2 model was trained with a 50 % chance of applying a linear combination to the acquired MRI, as opposed to 100 % for the LC-only model.

The RC+T1(Gd)+T2 model and the LC+T1(Gd)+T2 were statistically compared using image similarity metrics calculated per sequence. The best-performing model was chosen as the Domain Randomisation model for the final comparison, using MAE as the leading metric for model choice.

TABLE III
DETAILS FOR THE APPLICATION OF EARLY STOPPING FOR MODELS IN THE EXPERIMENTS.

Model	Sequences considered	Chosen iteration
Baseline	T1w, T1wGd, T2w	300,000
T1-only	T1w, T1wGd	400,000
RC+T1(Gd)	T1w, T1wGd	400,000
RC+T1(Gd)+T2	T1w, T1wGd, T2w	450,000
LC+T1(Gd)+T2	T1w, T1wGd, T2w	200,000

Early stopping was applied, as illustrated in G.

I. Final comparison

Three models were assessed in the final comparison: the best Domain Randomisation model (resulting from the choice of domain randomisation method in experiment III-H.3), the Baseline model (III-H.1), and the Baseline+FLAIR model.

The Baseline+FLAIR model was trained to obtain a measure for the best achievable performance for FLAIR input images. This model was trained with the same hyperparameters as before (section III-E) on the whole training set of 60 patients. The training dataset consisted of a mix of T1w (n = 30), T1wGd (n = 30), T2w images (n = 60) and FLAIR images (n = 60). Only the iteration at which to perform early stopping was finetuned, using MAE as the decision criterion while evaluating the sCT generated from the T1w, T1wGd, T2w and FLAIR images of the patients in the validation set. Early stopping was applied at iteration 450,000.

For the final comparison, the Baseline, Baseline+FLAIR and Domain Randomisation models were inferred on the test set (n = 25). Image similarity metrics and dosimetric accuracy were calculated for sCT generated by the three models from patients' T1w, T1wGd, T2w and FLAIR images in the test set. For each model, the image similarity metrics and metrics for dosimetric evaluation were statistically compared between the four sequences. Also, per sequence, statistical comparisons were made between the three models.

IV. RESULTS

The training time for the Baseline model was 32.0 h, while training the Baseline+FLAIR model took 46.2 h. For the Domain Randomisation model, the training time was 66.4 h. Inference time on the test set was approximately 4 s per sequence and patient for all models included in the final comparison.

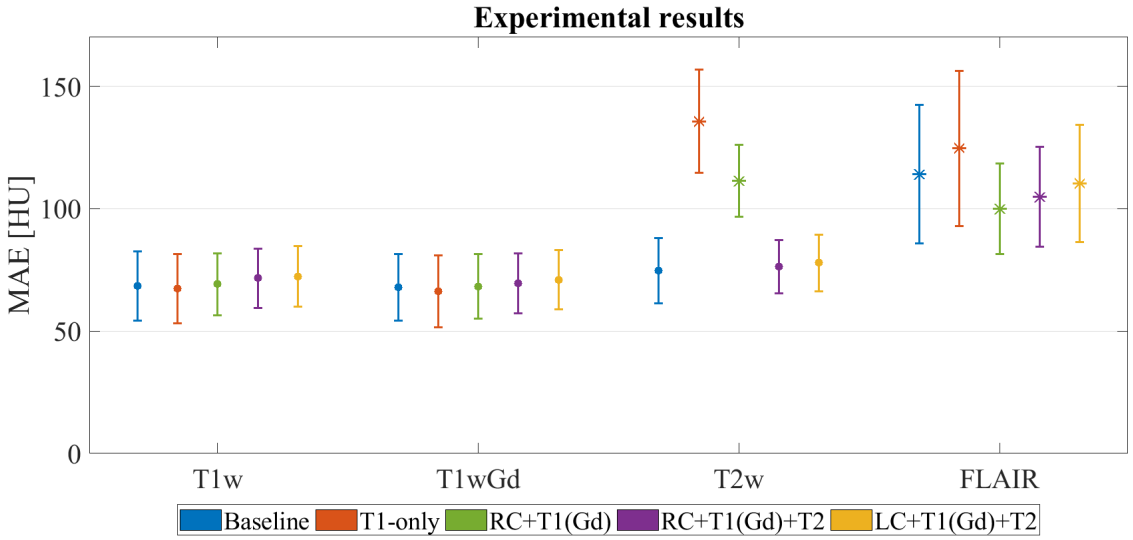

Fig. 7. Errorbar plot for the MAE in the intersection of the body contour of sCT compared to ground truth CT for the models presented in the experiments. Results were obtained on the validation set (n = 9) and are shown per sequence, from left to right: T1w, T1wGd, T2w and FLAIR. The marker indicates the mean value, with the errorbars representing the standard deviation. A dot marker indicates the sequence was used for model training, and an asterisk indicates the sequence was not part of the training data and only used for evaluation.

A. Experiments

1) Generalisation of conventional cGANs - Baseline vs T1-only model: For the Baseline model, the highest MAE was obtained on FLAIR images: 114 ± 28.4 HU (Fig. 7; table values including the range are reported in Appendix K). Removing T2w images from the training dataset of the T1-only model decreased the performance on T2w images compared to that obtained for the Baseline model. The performance of the T1-only model for T2w images was the worst among the sequences, resulting in an MAE of 136 ± 21.1 HU, while that obtained for the Baseline model was 74.7 ± 13.3 HU. Likewise, the MAE obtained for FLAIR images increased to 125 ± 31.6 HU. The increase in MAE for the T1-only model compared to the Baseline model was statistically significant for both sequences (p-values in Appendix L, Table XIII).

The slight decrease in MAE found for the T1-only model on T1w images (from 68.4 ± 14.2 HU to 67.2 ± 14.2 HU) was not statistically significant, in contrast to the decrease in MAE found on T1wGd images (from 67.8 ± 13.7 HU to 66.2 ± 14.7 HU). SSIM and PSNR (Appendix K) show the same trends in performance as MAE, although, for SSIM, the increase in T1w images for the T1-only model compared to the Baseline model was statistically significant, and the difference in PSNR and SSIM on T1wGd images was not (p-values in Appendix L, Table XIII).

2) Domain randomisation - Random contrast vs mixing sequences: Mixing T1w(Gd) images with RC images improved performance on T2w and FLAIR images compared to training only on T1w(Gd) images (Fig. 7, with table values in Appendix K). For T2w images, the MAE decreased from 136 ± 21.1 HU for the T1-only model to 111 ± 14.6 HU for the RC+T1(Gd) model ($p < 0.05$; Appendix L, Table XIII). Likewise, by adding RC images to the training data, the MAE for FLAIR images decreased from 125 ± 31.6 HU to 100 ± 18.5 HU ($p < 0.05$). For SSIM and PSNR, similar performance improvements were found (Appendix K). The slight performance decline on T1w(Gd) images was not statistically significant, except for a decrease in SSIM on T1w images.

Mixing T1w(Gd) images and RC images compared favourably to mixing T1w(Gd) images with T2w images (Baseline model) in terms of performance improvement on FLAIR compared to the T1-only model: the RC+T1(Gd) model resulted in an MAE of 100 ± 18.5 HU, compared to an MAE of 114 ± 28.4 HU for the Baseline model ($p < 0.05$). The exclusion of T2w images from the RC+T1w(Gd) model's training data significantly increased the MAE for T2w images compared to the Baseline model: 111 ± 14.6 HU vs 74.7 ± 13.3 HU. Performance differences between the Baseline and RC+T1(Gd) model on T1w(Gd) images were not statistically significant. The SSIM and PSNR were consistent with the MAE.

3) Domain randomisation - Random contrast vs linear combinations: For all sequences, the MAE was lower for the RC+T1(Gd)+T2 model than for the LC+T1(Gd)+T2 model (Fig. 7). Only the difference on FLAIR images was statistically significant (p-values in Appendix L, Table XIII): an MAE of 105 ± 20.5 HU was obtained for the RC+T1(Gd)+T2 model, compared to an MAE of 110 ± 23.9 HU for the LC+T1(Gd)+T2 model. For all sequences, the mean PSNR for the RC+T1(Gd)+T2 model was either higher than or equal to that obtained for the LC+T1(Gd)+T2 model (Appendix K). The same is true for SSIM, except for T1w images, where the SSIM was slightly higher for the LC+T1(Gd)+T2 model than for the RC+T1(Gd)+T2 model. Differences in SSIM and PSNR were not statistically significant. Overall, using RC images was deemed the most beneficial domain randomisation strategy of the two approaches. Consequently, the RC+T1(Gd)+T2 model was adopted as the Domain Randomisation model for final comparison.

B. Final comparison - Image similarity

1) Baseline model: The performance of the Baseline model on the test set (n = 25) was best on T1w and T1wGd images, with the difference between these two sequences not statistically significant for the three metrics (Fig. 8). The mean MAE was 64.2 ± 7.34 HU or 63.8 ± 9.12 HU for T1w and T1wGd images, respectively. The worst performance was found for FLAIR images, with a considerable difference with performance on T1w(Gd) images: the mean MAE was 106 ± 20.7 HU. Testing on T2w images resulted in a mean MAE of

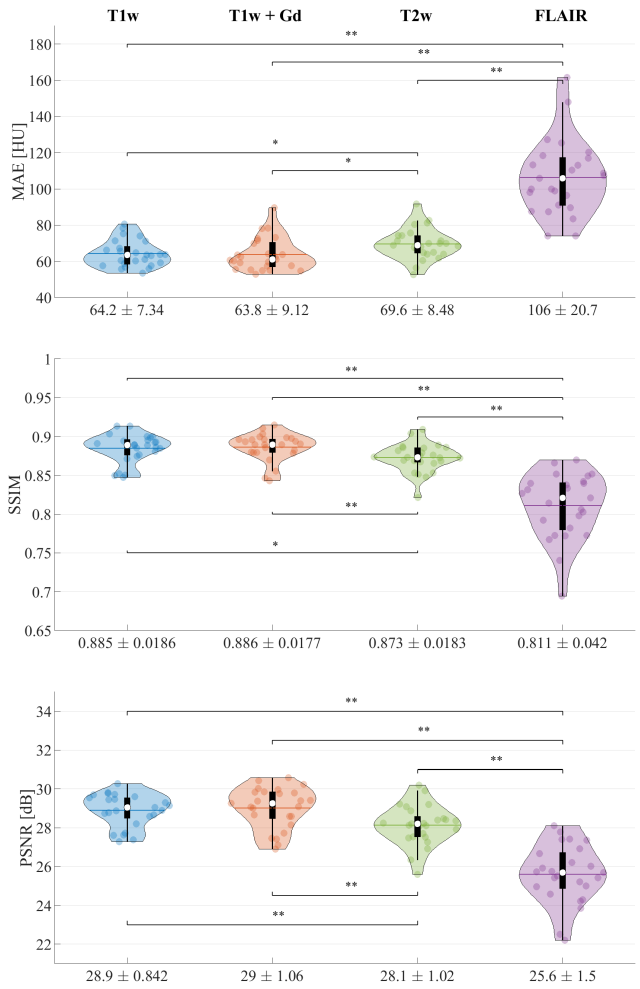

Fig. 8. Violin and box-and-whisker plots for the MAE (top), SSIM (middle) and PSNR (bottom) in the intersection of the body contour of sCT compared to ground truth CT for the Baseline model on the test set (n = 25). Results are presented per sequence, from left to right: T1w, T1wGd, T2w and FLAIR. The black box indicates the interquartile range and median (white circle) with whiskers indicating the range, outliers excluded. The width of the violin indicates the distribution of the data points. The mean values and standard deviations are shown. Statistically significant differences are indicated by * ($p < 0.05$) or ** ($p < 0.001$). Wilcoxon-signed rank tests were used for statistical testing.

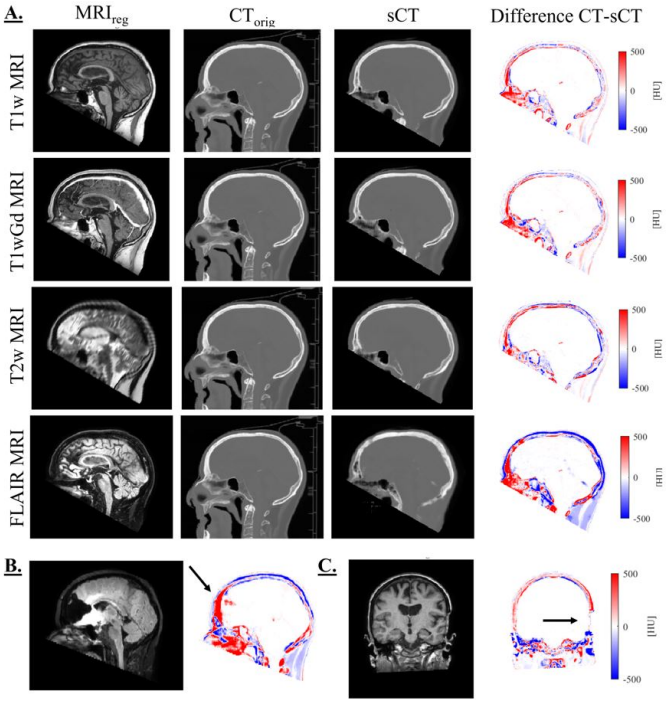

Fig. 9. Results produced by the Baseline model. A) Results for a representative subject for T1w, T1wGd, T2w and FLAIR input images (top to bottom). The image shows from left to right: original MRI, ground truth CT, sCT generated by the Baseline model, and the difference between the acquired CT and sCT. Typical problematic areas are the nasal cavities and the borders of the skull (bright in the image with the difference between CT and sCT on the right). For FLAIR specifically, the back of the neck (arrow) is problematic, and the skull is too thick on sCT, represented by the blue colour in the image with the difference between CT and sCT (right). B) Example patient with oedema in the frontal lobe. This area is hypointense on the FLAIR image (right), leading to problems in the sCT (arrow). C) Post-operative patient for whom part of the skull has been resected, which the Baseline model handles correctly (arrow). The result shown is for the T1w input image. Similar results are generated for the other sequences.

69.6 ± 8.48 HU. Results for SSIM and PSNR were in line with the MAE. For each metric, the difference in performance on FLAIR and T2w images compared to the performance on the respective other three sequences was statistically significant (Fig. 8).

Figure 9 shows results for three example cases. Typical problematic areas for all sequences are the skull border and the nasal cavities (Fig. 9 A). For FLAIR images specifically, the Baseline model produced sCTs on which the skull is thicker than on the acquired CT (Fig. 9 A). This finding translates to a bright blue colour in the image showing the difference between CT and sCT (right) due to the higher HU value assigned to the sCT than the acquired CT. Additionally, the back of the neck is typically a problematic area for FLAIR images (Fig. 9 A, arrow). Figure 9 B depicts the FLAIR image (left) and the corresponding image with the difference between CT and sCT (right) of a patient with an oedematous area in the frontal lobe. The area is hypointense on the FLAIR image, leading to an intensity similar to air in the sCT, which translates to a high positive value in the image with the difference between CT and sCT (right; bright red, arrow). Figure 9 C shows the T1w image of a patient with atypical anatomy: part of the skull has been resected for this post-surgical patient. Despite obtaining bottom performance for this patient for all sequences, the atypical anatomy was translated to the sCT satisfactorily by the Baseline model: the area of skull resection (arrow) shows a near-zero difference in HU values between ground truth CT and sCT. Similar results were obtained for this patient's other sequences (not shown). Figure 9 C also shows a substantial difference between CT and T1w-based sCT in the vertebrae, a common observation among the patients in the test set for each sequence.

2) Baseline+FLAIR model: As for the Baseline model, the best performance was found on T1w and T1wGd images for the Baseline+FLAIR model (Fig. 10). The difference in MAE and SSIM between these two sequences was not statistically significant, with a mean MAE of 65.5 ± 7.73 HU or 64.6 ± 9.25 HU for T1w and T1wGd images. Performance on T2w images was slightly worse with mean MAE = 71.2 ± 8.76 HU. All differences in the three metrics between T2w images and T1w and T1wGd metrics were statistically significant. The highest MAE was obtained for FLAIR images, with a mean value of 72.6 ± 10.1 HU. As for T2w images, differences between performance on FLAIR images and T1w and T1wGd images were statistically significant for all three metrics. Image similarity metrics on T2w images versus FLAIR images did not differ significantly.

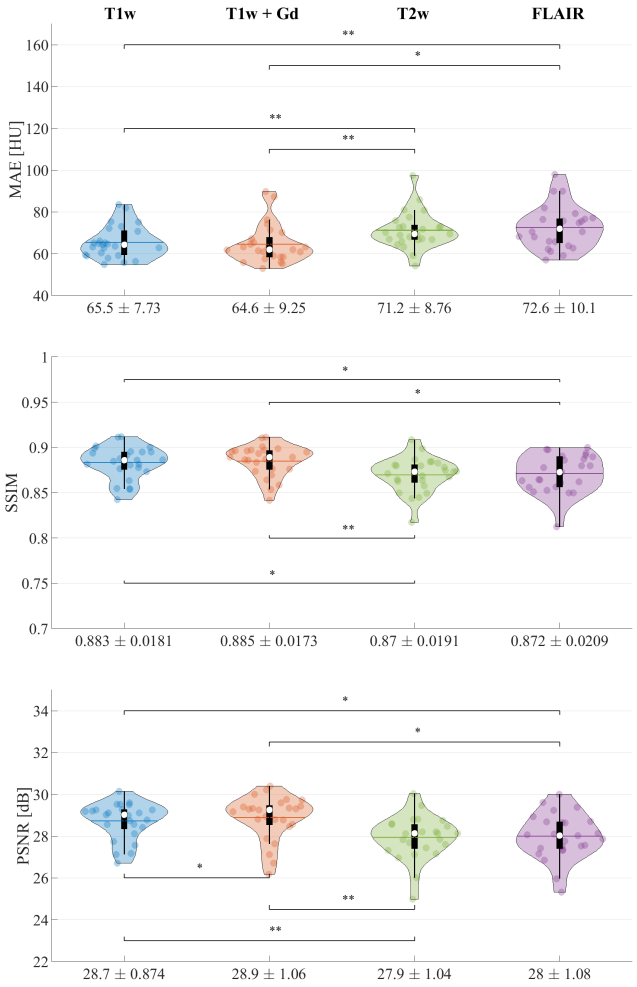

Fig. 10. Violin and box-and-whisker plots for the MAE (top), SSIM (middle) and PSNR (bottom) for the Baseline+FLAIR model on the test set (n = 25). The mean values and standard deviations are shown. Statistically significant differences are indicated by * ($p < 0.05$) or ** ($p < 0.001$). Significant improvements are seen in all three metrics for the FLAIR sequence (right) compared to the Baseline model.

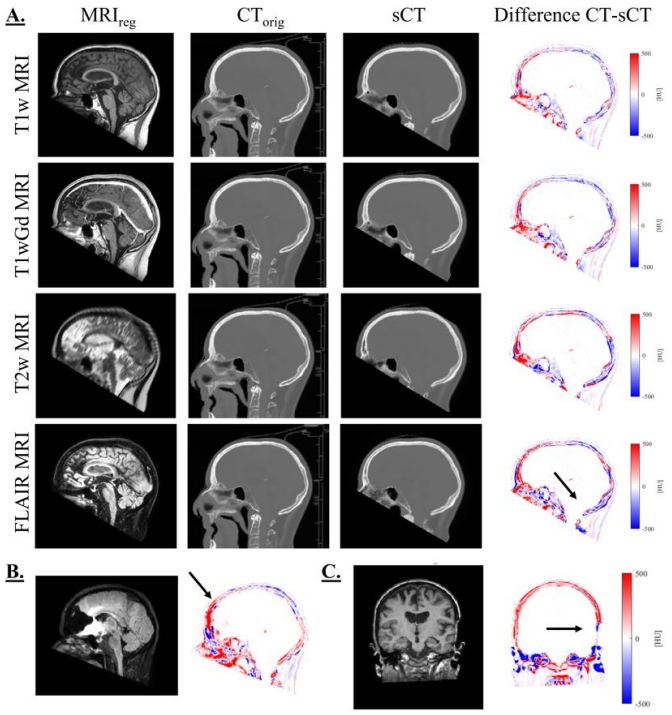

Fig. 11. Results generated by the Baseline+FLAIR model for the same example patients as presented for the Baseline model. A) Typical problematic areas in the T1w(Gd)- and T2w-based sCT are the same as for the Baseline model: nasal cavities and the borders of the skull (bright in the image with the difference between CT and sCT on the right). For FLAIR, the back of the neck (arrow) is less problematic than for the Baseline model, and the skull is no longer too thick on sCT. B) This patient's oedema in the frontal lobe (hypointense in the FLAIR image) leads to problems in the sCT (arrow), although less than for the Baseline model. C) The Baseline+FLAIR model handles this patient's partial skull resection decently (arrow).

Overall, the performance on T1w(Gd) and T2w images was slightly worse for the Baseline+FLAIR model than for the Baseline model, with an increased MAE and decreased PSNR and SSIM. For T2w images, this difference was statistically significant for all three metrics and for T1w images only for MAE and SSIM (p-values in Appendix L, Table XIV). Differences between the two models for T1wGd images were not statistically significant.

The most notable change in MAE was found on FLAIR images, favouring the Baseline+FLAIR model. Adding FLAIR images to the training data reduced the MAE from 106 ± 20.7 HU to 72.6 ± 10.1 HU. SSIM and PSNR results are in line with this finding ($p < 0.5$).

Visual inspection of sCTs generated by the Baseline+FLAIR model (Fig. 11) reveals specific problematic areas similar to those for the Baseline model (Fig. 9). Again, the skull border, the nasal cavities (Fig. 11 A), and the vertebrae (Fig. 11 C) prove difficult. While the Baseline model systematically produced sCTs from FLAIR images with the skull mapped too thick compared with the acquired CT (Fig. 9 A), this is not the case for the Baseline+FLAIR model (Fig. 11 A). Similarly, the problematic areas in the muscles of the neck observed for the Baseline model (Fig. 9 A, arrow) are less prominent for the Baseline+FLAIR model (Fig. 11 A, arrow). The hypointense oedematous area in the patient's frontal lobe (Fig. 11 B) is also closer to the acquired CT in the sCT generated by the Baseline+FLAIR model (arrow) than in the sCT generated by the Baseline model (Fig. 9 B, arrow). Still, the image shows a more considerable difference between acquired CT and sCT at the border between the skull and cerebrum near this hypointense region than at the rest of the border. In the sCT generated by the Baseline+FLAIR model for the post-surgical patient with partial skull resection (Fig. 11 C), the skull seems to continue further downwards than on the acquired CT, which translated to the blue colour in the image with the difference between CT and sCT (right image, arrow). Nevertheless, it is visible that part of the skull has been removed in this patient.

3) Domain Randomisation model: Similarly to what was found for the Baseline and Baseline+FLAIR models, the Domain Randomisation model performed best on T1w and T1wGd images (Fig. 12). An MAE of 67.6 ± 7.4 HU was obtained for T1w images and of 66.5 ± 9.2 HU for T1wGd images ($p > 0.05$). In line with the results obtained for the Baseline and Baseline+FLAIR model, the MAE found for T2w images was slightly higher: 71.5 ± 7.93 HU. The highest MAE was obtained for FLAIR images: 99.0 ± 14.9 HU. Differences in performance on T2w and FLAIR images compared to the other three sequences were all statistically significant. The SSIM and PSNR (Fig. 12, middle and bottom, respectively) are generally consistent with the findings for the MAE.

For the seen sequences (i.e., T1w, T1wGd and T2w images), the MAE obtained for the Domain Randomisation model was higher than that obtained for the Baseline and Baseline+FLAIR models,

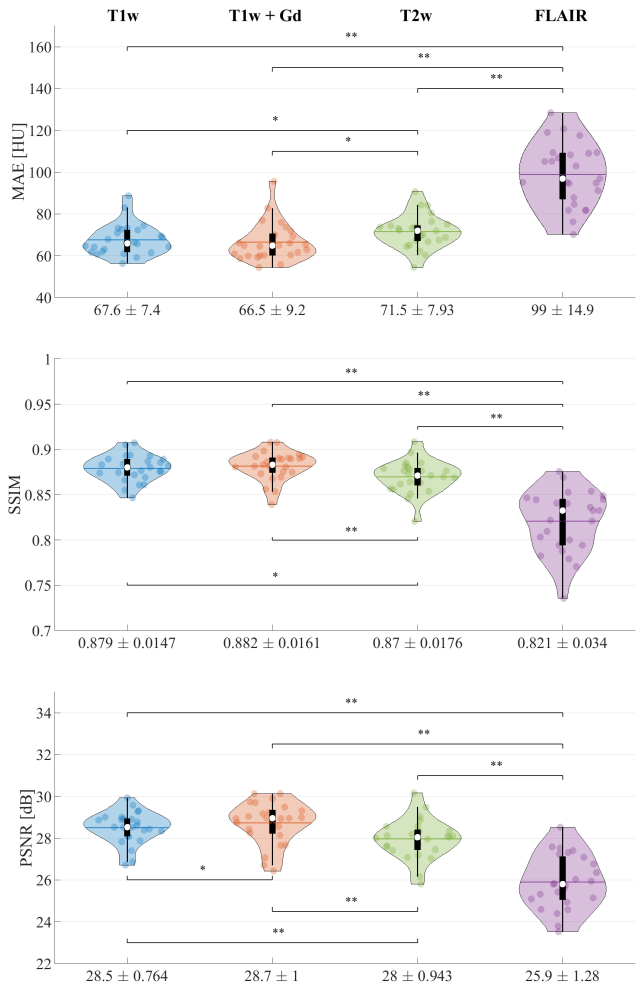

Fig. 12. Violin and box-and-whisker plots for the MAE (top), SSIM (middle) and PSNR (bottom) for the Baseline+FLAIR model on the test set (n = 25). The mean values and standard deviations are shown. Statistically significant differences are indicated by * ($p < 0.05$) or ** ($p < 0.001$). Significant improvements are seen in all three metrics for the FLAIR sequence (right) compared to the Baseline model, although performance on FLAIR does not reach the same level as that obtained for the Baseline+FLAIR model.

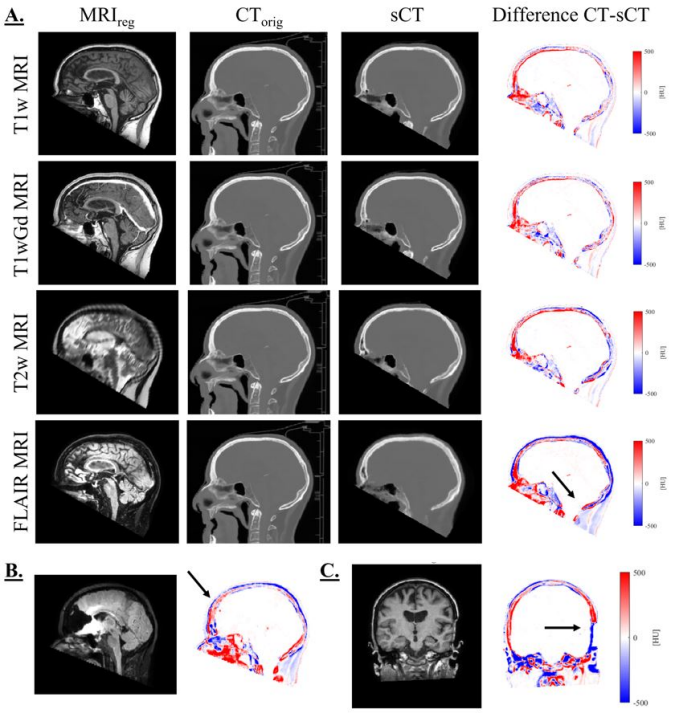

Fig. 13. Results generated by the Domain Randomisation model for the same example patients as presented for the Baseline and Baseline+FLAIR model. A) Typical problematic areas in the T1w(Gd)- and T2w-based sCT are the same as for the other two models: the nasal cavities and the borders of the skull (bright in the image with the difference between CT and sCT on the right). For FLAIR, the skull is too thick on sCT, like for the Baseline model. The back of the neck (arrow) is less problematic than for the Baseline model. B) This patient's oedema in the frontal lobe (hypointense in the FLAIR image) leads to problems in the sCT (arrow), although less than for the Baseline model. C) Post-operative patient for whom part of the skull has been removed, which is a problematic area for the Domain Randomisation model (arrow).

and the PSNR and SSIM were lower. All differences between the Domain Randomisation model and the Baseline model for these three sequences were statistically significant (p-values for comparisons between models in Table XIV, Appendix L). Likewise, the differences between the Domain Randomisation and the Baseline+FLAIR model were statistically significant for T1w and T1wGd images, but not for T2w images, with consistent results among the three metrics.

The MAE obtained for the Domain Randomisation model on FLAIR images was 7 HU lower than that obtained for the Baseline model ($p < 0.05$), a difference which is larger than the increase in MAE obtained for the other sequences (T1w: +3 HU, T1wGd: +3 HU; T2w: +2 HU). Despite this decrease in MAE on FLAIR images obtained through the addition of RC images during network training, the MAE obtained for the Domain Randomisation model was 26 HU higher than that achievable when adding FLAIR images to the training dataset (Baseline+FLAIR model; $p < 0.05$).

As for the Baseline and Baseline+FLAIR models, visual inspection of the generated sCTs reveals that the most prominent problematic areas for the Domain Randomisation model are the skull border and nasal cavities (Fig. 13 A) and the vertebrae (Fig. 13 C). While the neck muscles led to difficulties for the Baseline model (Fig. 9 A, arrow), this is partly resolved for the Domain Randomisation model (Fig. 13 A, arrow). Nevertheless, the image with the difference between CT and sCT (right) still shows more significant differences in this area than in the sCT generated by the Baseline+FLAIR model (Fig. 11 A, arrow). This reduction in differences between sCT and CT in muscle tissue compared to the Baseline model was observed for all patients in the test set. An unresolved problem in FLAIR-based sCTs is the mapping of the skull. Like the Baseline model (Fig. 9 A), the Domain Randomisation model systematically produced sCTs with the skull mapped thicker than on the acquired CT (Fig. 13 A).

The partial skull resection of the patient shown in Fig. 13 C is not mapped correctly by the Domain Randomisation model, unlike by the Baseline model (Fig. 9 C). In general, however, the Domain Randomisation model can handle abnormalities, as seen in Fig. 13 B. The mapping of the hypointensity in the frontal lobe visible in the FLAIR image (arrow) that led to problems for the Baseline model (Fig. 9 B, arrow) resembles the mapping obtained by the Baseline+FLAIR model (Fig. 11 B, arrow).

Overall, a visual inspection of the results for FLAIR images generated by the Domain Randomisation model reveals that the model might be more robust than the Baseline model. Figure 14 shows three example patients for whom the Baseline model produced artefacts in the sCT (green rectangles). Such artefacts were not observed in the FLAIR-based sCTs produced by the Baseline+FLAIR and Domain Randomisation models. The reduced difference between sCT and

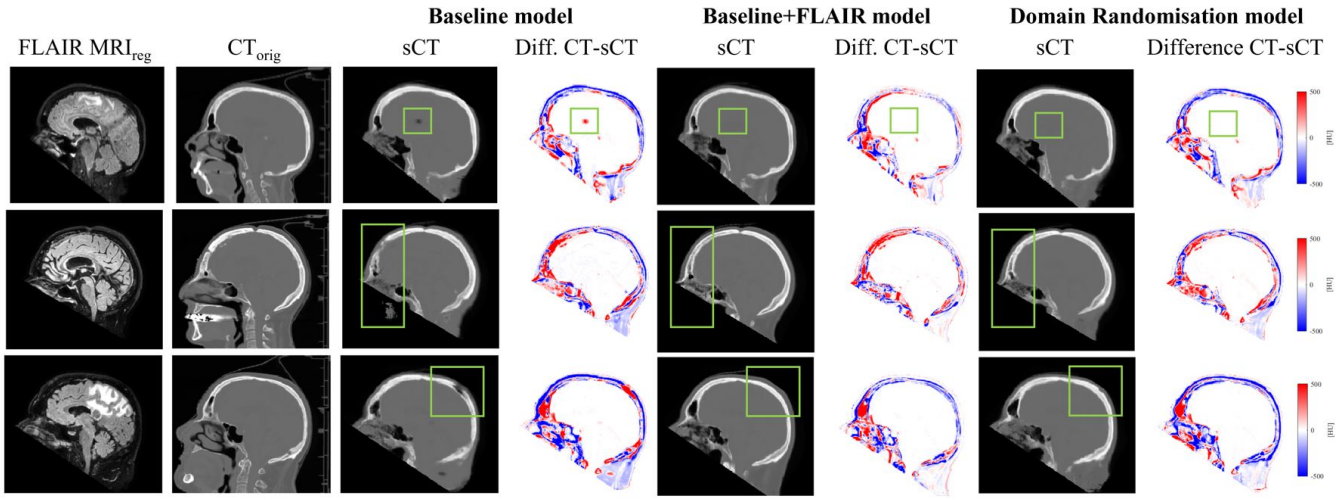

Fig. 14. Results from FLAIR images generated by the three models for three different patients. Images show from left to right: original FLAIR MRI, ground truth CT, sCT, and the difference between the acquired CT and sCT for the Baseline model, Baseline+FLAIR model and Domain Randomisation model, respectively. The areas marked with a green rectangle highlight artefacts in sCT produced by the Baseline model that are not present in the sCT generated by the Baseline+FLAIR and Domain Randomisation models.

CT in the muscles in the neck for sCT generated by the Domain Randomisation versus the Baseline model is also evident in Fig. 14.

C. Final comparison - Dosimetric accuracy

1) Baseline model: For the Baseline model, 3D γ-pass rates in the low dose region ($>$ 10 % of the prescribed dose) with 1%,1mm criterion were $>$ 95 % (Table IV) for every patient and each sequence. The pass rate obtained for FLAIR images (99.0 $\pm$ 1.1 %) was significantly lower than that computed for all other sequences (p-values in Appendix L, Table XV). The γ-pass rates with 3%,3mm and 2%,2mm criteria were $>$ 99 % for every patient and each sequence (Appendix M, Table XVII).

For the Baseline model, a DD in the high dose region ($>$ 90 % of the prescription dose) of -0.1 $\pm$ 0.2 % was obtained for treatment plans based on sCT generated from T1w, T1wGd and T2w images, and a DD of 0.4 $\pm$ 0.5 % was found for FLAIR images (Table IV). The DD was significantly larger for FLAIR than for the three other sequences (p-values in Appendix L, Table XV). Other differences in DD between sequences were not statistically significant.

For T1w-, T1wGd- and T2w-based sCT, the DD was below 1 % for every patient, while the DD in treatment plans from FLAIR-based sCT was $>$ 1 %, but $\leq$ 1.5 % for three patients (PT2, PT13 and PT18). There was a discrepancy between sCT and CT HU values for PT2 near the high dose region. The skull near the frontal lobe was imaged too thinly on sCT, causing HU values to be lower than in the acquired CT. Discontinuities were visible in the skull of this post-surgical patient in the problematic area, although no part of the skull was missing. For PT13 (Fig. 9 A), the high dose region was located in the dorsal part of the cerebrum, where differences in skull thickness occurred between the FLAIR-based sCT generated by the Baseline model and the acquired CT, this time with higher HU values in the sCT than in the acquired CT. Notable dose differences were observed for PT18 near the nasal cavities, close to one of the isocentres of irradiation. The sCT generated by the Baseline model from this patient's FLAIR image revealed more prominent differences in this area between HU values of sCT and acquired CT than the sCT generated for the other sequences.

In general, minor differences in D_{max} and D_{median} were observed for OARs in DVH analysis for each sequence (Appendix M). On average, differences were below 0.5 % for every sequence and DVH point. Individually, most patients had differences in DVH points $\leq$ 1 %. Exceptions were the lens (PT12) and the cochlea (PT5 and PT12) for T1w images, with differences $\leq$ 1.5 %. For T1wGd images, two patients had differences $\leq$ 1.5 % for the cochlea (PT5 and PT12). For T2w images, differences $\leq$ 2.0 % were found for one patient (PT12) in the lens, pituitary gland and cochlea. Differences in FLAIR images were $\leq$ 2.0 % for the pituitary gland (PT14), optic chiasm (PT1), and lens (PT12). PT12 patient had an RT plan with a vast irradiated area, matching the patient's large tumour volume. For this patient, for every MRI sequence, notable dose differences were observed around the body contour on the right half of the body.

2) Baseline+FLAIR model: 3D γ-pass rates in the low dose region with 1%,1mm criterion $>$ 97 % were obtained for each patient and MRI sequence (Table IV) for the Baseline+FLAIR model. As for the Baseline model, pass rates with 3%,3mm and 2%,2mm criteria were all $>$ 99 % (Appendix M, Table XVII). Statistically significant differences in pass rates between sequences were only found for the 1%,1mm criterion. For this criterion, the mean 3D γ-pass rate was 99.5 $\pm$ 0.7 % for T1w and T1wGd images, 99.4 $\pm$ 0.7 % for T2w images, and 99.4 $\pm$ 0.8 % for FLAIR images. Differences between T1w and T1wGd vs FLAIR images were statistically significant, as was the difference between T1w and T2w images (p-values in Appendix L, Table XV).

For FLAIR images, the Baseline+FLAIR model outperformed the Baseline model in terms of 3D γ-pass rates for the 1%,1mm criterion: 99.4 $\pm$ 0.8 % (Baseline+FLAIR model) vs 99.0 $\pm$ 1.1 % (Baseline model; p-values in Appendix L, Table XVI). Likewise, the other two pass rates were significantly higher for the Baseline+FLAIR model. Surprisingly, despite the higher MAE obtained in T1w images for the Baseline+FLAIR model versus the Baseline model, a significantly higher 3D γ-pass rate was obtained for the 1%,1mm criterion for the Baseline+FLAIR model (99.5 $\pm$ 0.7 % vs 99.4 $\pm$ 0.8 %). All other differences in pass rates between the two models were not significant.

For the Baseline+FLAIR model, differences in DD between sequences were not statistically significant (Table IV; p-values in Appendix L, Table XV). Absolute DD values obtained per sequence for the Baseline+FLAIR model were all smaller than those obtained for the Baseline model ($p < 0.05$; Appendix L, Table XVI).

The Baseline+FLAIR model generated sCTs resulting in treatment plans with a DD below 1.5 % for every patient and sequence. For

TABLE IV

DOSE EVALUATION ($\gamma_{1\%,1mm}$ AND DD) FOR SCT GENERATED BY THE BASELINE, BASELINE+FLAIR AND DOMAIN RANDOMISATION MODELS PER MRI SEQUENCE.

Metric	**Model**	*Sequence*			
		T1w	**T1wGd**	**T2w**	**FLAIR**
$\gamma_{1\%,1mm}$ [%] [a]	Baseline	99.4 ± 0.8 [97.1 - 100]	99.4 ± 0.8 [96.9 - 100]	99.4 ± 0.7 [97.3 - 100]	99.0 ± 1.1 [95.4 - 99.9]
	Baseline+FLAIR	99.5 ± 0.7 [97.4 - 100]	99.5 ± 0.7 [97.2 - 100]	99.4 ± 0.7 [97.4 - 100]	99.4 ± 0.8 [97.2 - 100]
	Domain Randomisation	99.4 ± 0.8 [97.0 - 100]	99.4 ± 0.8 [96.9 - 100]	99.3 ± 0.8 [97.2 - 100]	99.2 ± 0.9 [96.6 - 99.9]
DD [%] [b]	Baseline	-0.1 ± 0.2 [-0.5 - 0.1]	-0.1 ± 0.2 [-0.5 - 0.1]	-0.1 ± 0.2 [-0.4 - 0.8]	0.4 ± 0.6 [-1.0 - 1.5]
	Baseline+FLAIR	-0.02 ± 0.2 [-0.4 - 0.4]	-0.01 ± 0.2 [-0.7 - 0.5]	0.01 ± 0.3 [-0.4 - 1.1]	0.01 ± 0.4 [-1.4 - 0.7]
	Domain Randomisation	-0.1 ± 0.2 [0.5 - 0.2]	-0.2 ± 0.2 [0.5 - 0.1]	-0.1 ± 0.3 [-0.5 - 0.9]	0.3 ± 0.5 [-0.4 - 1.4]

Dosimetric accuracy was assessed through plan re-calculation on water-filled sCT compared to the water-filled acquired CT. Mean values and standard deviations ($\mu \pm 1\sigma$) and range ([min - max]) are reported. [a]Calculated in the D $>$ 10 % prescribed region. [b]Calculated in the D $>$ 90 % prescribed region.

T1w and T1wGd images, the DD was $<$ 1 % for every patient. For FLAIR images, the number of patients with a DD $>$ 1 % reduced to one (PT2) compared to three for the Baseline model. Similar to what was found for the Baseline model, for PT2, discrepancies between sCT and CT HU values were present near the high dose region in the area of surgical intervention. Unlike the Baseline model, for the Baseline+FLAIR model, there was one patient (PT16) for whom the DD was $>$ 1 % for T2w images. Partial volume effects were present on the border of the skull in this patient's T2w image, leading to substantial differences in HU values, with HU values in the sCT higher than those in the acquired CT. The area where partial volume effects occurred was located within the high dose region.

As for the Baseline model, differences in D_{max} and D_{median} were minor for all DVH points evaluated and all sequences, with average differences $<$ 0.5 % (boxplots in Appendix M). Most patients had differences in DVH points $\leq$ 1 % for every OAR, with some exceptions for T1wGd, T2w and FLAIR images. For T1wGd images, this was the case for one patient for the lens (PT12) and for another patient for the brainstem (PT18), with both differences $\leq$ 1.5 %. The lens and cochlea (both PT12) and the brainstem (PT18) were exceptions for T2w images, with differences $\leq$ 2.0 % for all. Finally, for FLAIR images, one patient (PT12) had a difference of $\leq$ 1.5 % in the cochlea. As observed for the Baseline model, sCT-based dose plans for PT12 resulted in notable dose differences at the body contour on the right side of the body.

3) Domain Randomisation model: The 3D γ-pass rates in the low dose region with 1%,1mm criterion obtained for the Domain Randomisation model were $>$ 96 % for each patient and each MRI sequence (Table IV). For the 3%,3mm and 2%,2mm criteria, pass rates were all $>$ 99 %, as for the other two models (Appendix M, Table XVII). As for the Baseline model, for the 1%,1mm criterion, the pass rate obtained for FLAIR images (99.2 ± 0.9 %) was lower than the pass rates obtained for all other sequences, with $p < 0.05$ (p-values in Appendix L, Table XV).

Differences in 3D γ-pass rates between the Baseline and Domain Randomisation model were insignificant for the seen sequences (Appendix L, Table XVI). However, for FLAIR images, the Domain Randomisation model outperformed the Baseline model for the 1%,1mm criterion (99.2 ± 0.9 % vs 99.0 ± 1.1 %; $p < 0.05$).

Compared to the Baseline+FLAIR model, for FLAIR images, the Domain Randomisation model resulted in significantly lower 3D γ-pass rates for the 1%,1mm criterion (99.2 ± 0.9 % vs 99.4 ± 0.8 %). The Domain Randomisation model also resulted in significantly lower pass rates for FLAIR for the 3%,3mm, but not the 2%,2mm criterion. Additionally, higher pass rates were obtained for the Baseline+FLAIR model than for the Domain Randomisation model for T1w images for the 1%,1mm and 3%,3mm criteria ($p < 0.05$).

As for the Baseline model, for the Domain Randomisation model, the DD in the high dose region ($>$ 90 % of the prescription dose) obtained for FLAIR images (0.3 ± 0.5 %; Table IV) was significantly higher than that obtained for the other three sequences; see the corresponding p-values in Appendix L, Table XV. Differences in DD between the Domain Randomisation model and the Baseline model were not significant (Appendix L, Table XVI). Comparing the DD obtained for the Domain Randomisation and Baseline+FLAIR models resulted in p-values $<$ 0.05 for every sequence, with the DD obtained for the Baseline+FLAIR model smaller in absolute terms.

For the Domain Randomisation model, the DD was below 1 % for every patient for T1w(Gd)- and T2w-based sCT. Meanwhile, the DD in treatment plans from FLAIR-based sCT was $>$ 1 %, but $<$ 1.5 % for three patients (PT13, PT16 and PT18). As for the Baseline model, PT13's sCT generated by the Domain Randomisation model showed an overestimated skull thickness near the high dose region near the dorsal cerebrum. For PT18, areas with notable dose differences were similar to those observed for the Baseline model, mainly the nasal cavities. For PT16, dose differences in the high dose region were substantial along the inner border of the skull, in line with the general observation that the MAE along the skull border was comparatively high for sCT generated from FLAIR images. The area of high DD was less localised and more spread out over the skull border than the area of high DD observed for this patient for the dose plan from the T2w-based sCT generated by the Baseline+FLAIR model, which was attributed to partial volume effects that were not observed here.

In general, boxplots for differences in DVH points for OARs (Appendix M) reveal slight differences in D_{max} and D_{median} for all DVH points and sequences, with average differences $<$ 0.5 % as for the other two models. On an individual basis, most patients had differences in DVH points $\leq$ 1 % for every OAR. Exceptions for T1w images were one patient that had a difference of $<$ 1.5 % for the cochlea (PT6) and another for the brainstem (PT18). For T2w images, one patient (PT12) had differences $<$ 1.5 % for the lens, cochlea and pituitary gland. FLAIR images resulted in DVH

differences < 2.5 % for the pituitary gland and lens for the same patient. A DVH difference < 1.5 % was found for another patient (PT6) for FLAIR images for the cochlea. The patient with the most considerable differences in DVH points (PT12) was the same as for the other two models. As for the other models, dose differences were substantial along the body contour on the right side of the body.

V. Discussion

In this work, we investigated the influence of domain randomisation on the problem of MRI-to-sCT generation, investigating its impact on rendering a cGAN contrast-agnostic. To the best of our knowledge, this is the first work[2] exploring whether a single cGAN network can be trained for sCT generation from various MRI sequences without the need for (re)training the network on unseen sequences (here: FLAIR).

A. Overall model performance

Image similarity of our three models from the final comparison on their seen sequences is on par with that obtained in other work (Appendix O). Our highest MAE for T1w(Gd) images, obtained for the Domain Randomisation model, falls within the range reported in the literature (T1w: 45.4 HU $\pm$ 8.52 HU [83] to 131 $\pm$ 14.3 HU [84]; T1wGd: 44.6 $\pm$ 7.48 HU [83] to 89.3 $\pm$ 10.3 HU [85]). Likewise, the performance of our Baseline+FLAIR model for FLAIR falls within the range reported in the limited number of studies using T2w FLAIR images for model training: 51.2 $\pm$ 4.5 HU [83] to 115 $\pm$ 22 HU [86]. The slightly higher MAE obtained for our Baseline+FLAIR model on this sequence compared to T1w(Gd) is in line with findings in [83, 86, 87] and is attributed to the contrast in FLAIR images. Specific tissues, e.g., muscle, are generally low in image intensity, and the skull border is more difficult to distinguish by eye than in the other sequences. We obtained slightly higher mean MAE values for T2w than earlier work (maximum: 68.3 $\pm$ 7.3 HU [57]). Our T2w images were acquired with a larger slice thickness than the other sequences, leading to partial volume effects (Appendix P) and explaining the slightly reduced performance compared to T1w(Gd). Mean γ-pass rates were on the high end of values in the literature (Appendix O). Altogether, also considering that the number of patients included in our training set (n = 60) is significantly larger than the median number in other studies (n = 33), we believe that our models' performance for the seen sequences is sufficient to explore generalisation.

Nevertheless, some limitations should be taken into account. A supervised framework was adopted, requiring a set of well-registered MRI-CT pairs. Poor registration between the CT and MRI in the training dataset has been shown to negatively influence DL network performance for MRI-to-CT synthesis when employing supervised training [88]. Additionally, registration imperfections in the test set might lead to an underestimated network performance (Appendix P). In this work, the outline of the skull on sCT showed imperfections for all sequences, and sinuses and vertebrae were especially problematic. We expect that the errors in rigid registration are also the most pronounced in these locations. In the future, non-rigid registrations could be explored to improve the overall model performance. An alternative approach that might alleviate problems associated with image registration is unpaired training [41].

Hyperparameter tuning was performed using only T1w images. A slightly different hyperparameter combination might improve performances when mixing sequences or adding RC images. A small check of the Domain Randomisation model's hyperparameters revealed that this might be true, with the current combination leading to 2-3 HU higher MAE on T1w(Gd) and T2w than the best combination tested. However, to isolate the effect of domain randomisation, the three final models were trained using the same hyperparameter set. Future work should clarify whether performance can be improved through hyperparameter finetuning. Based on the findings of the hyperparameter check, we expect the maximum obtainable performance improvement on the seen sequences to be in the order of 5 HU.

A critical note should be made about our dosimetric evaluation. Differences in the FOV of the acquired MRI and planning CT led to equal differences between the sCT and planning CT. Water-filling was used to avoid the calculated dosimetric accuracy being mostly a measure of this difference in FOVs. For beams in the treatment plan passing through the water-filled area, the dosimetric accuracy of the sCT could be overestimated. To cope with this limitation, we have adopted stricter criteria for clinical acceptability of dose plans than generally employed in practice (section III-C.2). The sCT developed in this work should not be considered directly for clinical use but are still relevant to shed light on model generalisation, which is this study's primary aim. In the future, models should be trained using MRI data covering a larger FOV (e.g., up to the chin) to evaluate the dosimetric accuracy without water-filling.

B. Generalisation and domain randomisation

We found that a model trained on T1w(Gd) images only (section IV-A.1) generalised poorly to FLAIR and T2w images, in line with the results for generalisation to unseen sequences in [56, 57]. Mixing T1w(Gd) and T2w images in the training data (Baseline model) improved image similarity metrics, not only for T2w but also for FLAIR images, although the performance for FLAIR was still inferior to that for the other sequences. The accuracy of dose plans generated from sCT was generally high for all models and sequences. Considering the adopted criteria for clinical acceptability of 3D γ-pass rates with 3%,3mm and 2%,2mm criteria, dose plans were acceptable for all patients and sequences, even for FLAIR-based sCT generated by the Baseline model. Therefore, we considered the more stringent 1%,1mm criterion for further comparisons.

We compared two approaches for domain randomisation, proving that the method involving synthetic RC images was more effective than the method based on linear combinations of acquired T1w(Gd) and T2w images. The Domain Randomisation model generated sCT from FLAIR images with significantly improved image similarity and dosimetric accuracy compared to the Baseline model.

We found that the benefit of adding RC images to the training data was more substantial when only T1w acquired MRI were used for network training (RC+T1(Gd) vs T1-only model) than when both T1w(Gd) and T2w images were used (RC+T1(Gd)+T2 vs Baseline model). This finding is in line with the finding that mixing T1w(Gd) images and T2w images already improved generalisation to FLAIR compared to training on T1w(Gd) images only. Perhaps mixing sequences provides the network with similar clues about focusing on contrast-invariant features as adding RC images.

A substantial gap remained in performance on FLAIR images for the Domain Randomisation model compared to the Baseline+FLAIR model regarding both image similarity and dose accuracy. Generating RC images from label maps results in a loss of within-label structure and detail compared to the acquired MRI. We hypothesise that this loss of detail reduced the information in (part of) the network's training data needed for a per-voxel mapping to a CT representation. Possibly, such an information reduction could explain why the Domain Randomisation model's performance for FLAIR did not reach the same level as obtained when adding FLAIR to the training

[2]The search strategy adopted to verify that no other work provided a similar exploration is provided in Appendix N.

data. Future research should investigate whether model performance on unseen sequences can be pushed further toward the performance achieved when training on this sequence.

Due to the different tasks (segmentation vs synthesis) in this work and Billot *et al.* [58, 59], a quantitative comparison of the effect of the applied domain randomisation method is difficult. Nevertheless, while Billot *et al.* [58] state their segmentation network is accurate across all test domains, our network for sCT synthesis does not perform equally well on FLAIR as on T1w(Gd) or T2w images from held-out test patients. Note that in [59], the segmentation network was also tested on FLAIR images and obtained a lower performance than on various T1w and T2w datasets. However, Billot *et al.* [59] did not directly compare, with a statistical test, the performance on FLAIR to the performance on other sequences, which complicates judging to what extent their model is contrast-agnostic.

A limitation in this work is that our segmentations were not, per definition, perfectly aligned with the corresponding ground truth (acquired CT), unlike in [58, 59] where segmentations served as the basis for generating training data and as the ground truth. This drawback might have reduced the effect of the RC images on network performance in our work. Also, visual inspection of the label maps showed imperfections: the eyes and lenses were frequently smaller than in the acquired MRI. Additionally, the label for the brain stem did not extend to the caudal end of the MRI FOV. Moreover, the final segmentation was partly derived from the patients' T1w MRI and partly from their CT (i.e., the bone and soft tissue labels). Obtaining the best possible segmentations, e.g. through manual segmentation, is expensive and was out of scope for this research. Future studies should clarify whether more accurate or elaborate label maps are more suitable as the basis for RC images.

The domain randomisation method using RC images requires segmenting (extra)cerebral structures. These might not always be available when training a model for MRI-to-CT translation. Therefore, this method might not always be practical for sCT synthesis.

A second domain randomisation method, based on linear combinations of acquired T1w(Gd) and T2w images, was tested. To the best of our knowledge, this is the first work investigating such a method for sCT generation. An advantage of this method over RC images is that it requires minimal effort and is easily applicable if multiple sequences are available per patient. However, the method is more constrained and the variability of the generated training data is much more limited, which could explain why this method is not as effective as using RC images. Theory and earlier studies suggest that variability beyond what the network will encounter in reality can be beneficial [55, 60, 89], in line with findings in [59], where synthetic images mimicking specific MRI sequences proved counterproductive.

C. Outlook and clinical implications

We limited our investigation to RC and LC as methods for domain randomisation. Future work could explore other methods, like extending LC to non-linear combinations or increasing the number of acquired MRI sequences used for combination. Also, the variability in the RC images could be further increased, e.g., using random elastic deformations or simulation of bias field artefacts as in [58, 59]. Alternatively, one could combine RC and LC, e.g., by overlaying specific labels over an LC image or using label maps to create label-specific offsets in the voxel values of acquired MRI. Another approach could explore using GANs or other DL models to generate synthetic training data, as suggested in, e.g., [90, 91].

An open question is whether the implemented domain randomisation approach could already be employed clinically to bridge smaller domain gaps than an entirely new sequence, like same-sequence data from a different hospital or changes in the acquisition protocol that might occur over time. Further evaluations on new datasets are needed to investigate whether this is the case.

This work provides the first attempt toward contrast-agnostic sCT generation for MR-only RT planning. A clear improvement was found in image similarity for sCT generated from an unseen sequence by our Domain Randomisation model compared to a Baseline model. A slight deterioration was allowed in image similarity for the seen sequences to achieve such improvement. Interestingly, in terms of dosimetric accuracy, our Baseline model already achieved good results for most patients for the unseen sequence simply by training on a mix of other sequences. The Domain Randomisation model improved the 3D γ-pass rate with 1%,1mm criterion for this unseen sequence. In contrast, differences with the Baseline model in dose metrics were not statistically significant for the seen sequences, leading us to believe that the small decrease in image similarity obtained for the seen sequences is clinically acceptable. Moreover, the Domain Randomisation model reduced artefacts observed in FLAIR-based sCT compared to the Baseline model. Altogether, the results indicate that domain randomisation can improve generalisation to unseen sequences for sCT generation. Before clinically implementing the methods described in this work, dosimetric accuracy must be evaluated in a clinical setting on MRI acquired with a larger FOV.

The results obtained in this work indicate that domain randomisation might help reduce the need for network re-training if the model is to be used on a sequence unseen during network training. This would be helpful if exceptions need to be made in imaging protocols for specific patients, e.g., due to allergies or claustrophobia. On the other hand, we are unsure whether the performance improvement found in this work is substantial enough to justify the effort associated with obtaining segmentations, if not already available. Therefore, we advise developing the method further before clinical implementation to either a) push performance on unseen sequences further towards the performance achieved when including the sequence in the training data, or b) design a simplified method that obviates the need for label maps, e.g., by extending the LC method.

VI. Conclusion

We investigated the ability to generalise to unseen sequences of a DL model tasked with sCT synthesis for MR-only radiotherapy. We found that mixing acquired sequences in the training data improved image similarity for an unseen sequence compared to only training on a single sequence. We considered two methods for domain randomisation, showing that adding random contrast images generated from label maps to the training data is more effective than applying random linear combinations of acquired MRI. Before clinical implementation, the domain randomisation method should be developed further to test whether performance on unseen sequences can be obtained closer to that achieved for seen sequences.

Generally, a satisfactory dosimetric accuracy was obtained when training on a mix of acquired sequences, even for the unseen sequence. However, the adopted domain randomisation method improved dosimetric accuracy and image similarity on this unseen sequence, indicating that domain randomisation could help reduce the need for network re-training if the model is to be used on a sequence unseen during network training.

References

[1] M. B. Barton, S. Jacob, J. Shafiq, K. Wong, S. R. Thompson, T. P. Hanna *et al.*, "Estimating the demand for radiotherapy from the evidence: A review of changes from 2003 to 2012," *Radiotherapy and Oncology*, vol. 112, no. 1, pp. 140–144, 2014. [Online]. Available: https://doi.org/10.1016/j.radonc.2014.03.024

[2] J. Seco and P. M. Evans, "Assessing the effect of electron density in photon dose calculations," *Medical Physics*, vol. 33, no. 2, pp. 540–552, 2006. [Online]. Available: https://doi.org/10.1118/1.2161407

[3] C. F. Njeh, "Tumor delineation: The weakest link in the search for accuracy in radiotherapy," *Journal of Medical Physics*, vol. 33, no. 4, pp. 136–140, 2008. [Online]. Available: https://doi.org/10.4103/0971-6203.44472

[4] P. Dirix, K. Haustermans, and V. Vandecaveye, "The Value of Magnetic Resonance Imaging for Radiotherapy Planning," *Seminars in Radiation Oncology*, vol. 24, no. 3, pp. 151–159, 2014. [Online]. Available: https://doi.org/10.1016/j.semradonc.2014.02.003

[5] M. Jolicoeur, M.-L. Racine, I. Trop, L. Hathout, D. Nguyen, T. Derashodian *et al.*, "Localization of the surgical bed using supine magnetic resonance and computed tomography scan fusion for planification of breast interstitial brachytherapy," *Radiotherapy and Oncology*, vol. 100, no. 3, pp. 480–484, 2011. [Online]. Available: https://doi.org/10.1016/j.radonc.2011.08.024

[6] C. Rasch, R. Steenbakkers, and M. Van Herk, "Target definition in prostate, head, and neck," *Seminars in Radiation Oncology*, vol. 15, no. 3, pp. 136–145, 2005. [Online]. Available: https://doi.org/10.1016/j.semradonc.2005.01.005

[7] M. Just, H. P. Rösler, H. P. Higer, J. Kutzner, and M. Thelen, "MRI-assisted radiation therapy planning of brain tumors-clinical experiences in 17 patients," *Magnetic Resonance Imaging*, vol. 9, no. 2, pp. 173–177, 1991. [Online]. Available: https://doi.org/10.1016/0730-725X(91)90007-9

[8] N. Datta, R. David, R. Gupta, and P. Lal, "Implications of contrast-enhanced CT-based and MRI-based target volume delineations in radiotherapy treatment planning for brain tumors," *Journal of Cancer Research and Therapeutics*, vol. 4, no. 1, pp. 9–13, 2008. [Online]. Available: https://doi.org/10.4103/0973-1482.39598

[9] J. H. Jonsson, M. G. Karlsson, M. Karlsson, and T. Nyholm, "Treatment planning using MRI data: an analysis of the dose calculation accuracy for different treatment regions," *Radiation Oncology*, vol. 5, no. 1, p. 62, 2010. [Online]. Available: https://doi.org/10.1186/1748-717X-5-62

[10] G. Brix, H. Kolem, W. R. Nitz, M. Bock, A. Huppertz, C. J. Zech *et al.*, "Basics of Magnetic Resonance Imaging and Magnetic Resonance Spectroscopy," in *Magnetic Resonance Tomography*, M. F. Reiser, W. Semmler, and H. Hricak, Eds. Berlin, Germany: Springer Berlin Heidelberg, 2008, ch. 2, pp. 3–167. [Online]. Available: https://doi.org/10.1007/978-3-540-29355-2_2

[11] M. A. Schmidt and G. S. Payne, "Radiotherapy planning using MRI," *Physics in Medicine and Biology*, vol. 60, no. 22, pp. R323–R361, 11 2015. [Online]. Available: https://doi.org/10.1088/0031-9155/60/22/R323

[12] K. Ulin, M. M. Urie, and J. M. Cherlow, "Results of a multi-institutional benchmark test for cranial CT/MR image registration," *International Journal of Radiation Oncology*Biology*Physics*, vol. 77, no. 5, pp. 1584–1589, 2010. [Online]. Available: https://doi.org/10.1016/j.ijrobp.2009.10.017

[13] P. L. Roberson, P. W. McLaughlin, V. Narayana, S. Troyer, G. V. Hixson, and M. L. Kessler, "Use and uncertainties of mutual information for computed tomography/magnetic resonance (CT/MR) registration post permanent implant of the prostate," *Medical Physics*, vol. 32, no. 2, pp. 473–482, 2005. [Online]. Available: https://doi.org/10.1118/1.1851920

[14] T. Nyholm, M. Nyberg, M. G. Karlsson, and M. Karlsson, "Systematisation of spatial uncertainties for comparison between a MR and a CT-based radiotherapy workflow for prostate treatments," *Radiation Oncology*, vol. 4, no. 1, p. 54, 2009. [Online]. Available: https://doi.org/10.1186/1748-717X-4-54

[15] Y. K. Lee, M. Bollet, G. Charles-Edwards, M. A. Flower, M. O. Leach, H. McNair *et al.*, "Radiotherapy treatment planning of prostate cancer using magnetic resonance imaging alone," *Radiotherapy and Oncology*, vol. 66, no. 2, pp. 203–216, 2003. [Online]. Available: https://doi.org/10.1016/S0167-8140(02)00440-1

[16] T. Nyholm and J. Jonsson, "Counterpoint: Opportunities and Challenges of a Magnetic Resonance Imaging–Only Radiotherapy Work Flow," *Seminars in Radiation Oncology*, vol. 24, no. 3, pp. 175–180, 2014. [Online]. Available: https://doi.org/10.1016/j.semradonc.2014.02.005

[17] M. Kapanen, J. Collan, A. Beule, T. Seppälä, K. Saarilahti, and M. Tenhunen, "Commissioning of MRI-only based treatment planning procedure for external beam radiotherapy of prostate," *Magnetic Resonance in Medicine*, vol. 70, no. 1, pp. 127–135, 2013. [Online]. Available: https://doi.org/10.1002/mrm.24459

[18] P. Khong, H. Ringertz, V. Donoghue, D. Frush, M. Rehani, K. Appelgate *et al.*, "ICRP Publication 121: Radiological Protection in Paediatric Diagnostic and Interventional Radiology," *Annals of the ICRP*, vol. 42, pp. 1–63, 4 2013. [Online]. Available: https://doi.org/10.1016/j.icrp.2012.10.001

[19] J. M. Edmund and T. Nyholm, "A review of substitute CT generation for MRI-only radiation therapy," *Radiation Oncology*, vol. 12, no. 1, p. 28, 2017. [Online]. Available: https://doi.org/10.1186/s13014-016-0747-y

[20] A. M. Owrangi, P. B. Greer, and C. K. Glide-Hurst, "MRI-only treatment planning: Benefits and challenges," *Physics in Medicine and Biology*, vol. 63, no. 5, 2018. [Online]. Available: https://doi.org/10.1088/1361-6560/aaaca4

[21] M. Karlsson, M. G. Karlsson, T. Nyholm, C. Amies, and B. Zackrisson, "Dedicated Magnetic Resonance Imaging in the Radiotherapy Clinic," *International Journal of Radiation Oncology*Biology*Physics*, vol. 74, no. 2, pp. 644–651, 2009. [Online]. Available: https://doi.org/10.1016/j.ijrobp.2009.01.065

[22] S. Devic, "MRI simulation for radiotherapy treatment planning," *Medical Physics*, vol. 39, no. 11, pp. 6701–6711, 2012. [Online]. Available: https://doi.org/10.1118/1.4758068

[23] J. J. W. Lagendijk, B. W. Raaymakers, A. J. E. Raaijmakers, J. Overweg, K. J. Brown, E. M. Kerkhof *et al.*, "MRI/linac integration," *Radiotherapy and Oncology*, vol. 86, no. 1, pp. 25–29, 2008. [Online]. Available: https://doi.org/10.1016/j.radonc.2007.10.034

[24] S. Mutic and J. F. Dempsey, "The ViewRay System: Magnetic Resonance–Guided and Controlled Radiotherapy," *Seminars in Radiation Oncology*, vol. 24, no. 3, pp. 196–199, 2014. [Online]. Available: https://doi.org/10.1016/j.semradonc.2014.02.008

[25] W. A. Hall, E. S. Paulson, U. A. van der Heide, C. D. Fuller, B. W. Raaymakers, J. J. W. Lagendijk *et al.*, "The transformation of radiation oncology using real-time magnetic resonance guidance: A review," *European Journal of Cancer*, vol. 122, pp. 42–52, 2019. [Online]. Available: https://doi.org/10.1016/j.ejca.2019.07.021

[26] B. W. Raaymakers, I. M. Jürgenliemk-Schulz, G. H. Bol, M. Glitzner, A. N. T. J. Kotte, B. van Asselen *et al.*, "First patients treated with a 1.5 T MRI-Linac: clinical proof of concept of a high-precision, high-field MRI guided radiotherapy treatment," *Physics in Medicine and Biology*, vol. 62, no. 23, pp. L41–L50, 2017. [Online]. Available: https://doi.org/10.1088/1361-6560/aa9517

[27] M. F. Spadea, M. Maspero, P. Zaffino, and J. Seco, "Deep learning-based synthetic-CT generation in radiotherapy and PET: a review," *Medical Physics*, vol. 48, no. 11, pp. 6537–6566, 2021. [Online]. Available: https://doi.org/10.1002/mp.15150

[28] L. G. W. Kerkmeijer, M. Maspero, G. J. Meijer, J. R. N. van der Voort van Zyp, H. C. J. de Boer, and C. A. T. van den Berg, "Magnetic Resonance Imaging only Workflow for Radiotherapy Simulation and Planning in Prostate Cancer," *Clinical Oncology*, vol. 30, no. 11, pp. 692–701, 2018. [Online]. Available: https://doi.org/10.1016/j.clon.2018.08.009

[29] J. Sjölund, D. Forsberg, M. Andersson, and H. Knutsson, "Generating patient specific pseudo-ct of the head from mr using atlas-based regression," *Physics in Medicine and Biology*, vol. 60, no. 2, pp. 825–839, 2015. [Online]. Available: https://doi.org/10.1088/0031-9155/60/2/825

[30] E. Johnstone, J. J. Wyatt, A. M. Henry, S. C. Short, D. Sebag-Montefiore, L. Murray *et al.*, "Systematic Review of Synthetic Computed Tomography Generation Methodologies for Use in Magnetic Resonance Imaging–Only Radiation Therapy," *International Journal of Radiation Oncology*Biology*Physics*, vol. 100, no. 1, pp. 199–217, 2018. [Online]. Available: https://doi.org/10.1016/j.ijrobp.2017.08.043

[31] M. F. Spadea, G. Pileggi, P. Zaffino, P. Salome, C. Catana, D. Izquierdo-Garcia *et al.*, "Deep Convolution Neural Network (DCNN) Multiplane Approach to Synthetic CT Generation From MR images—Application in Brain Proton Therapy," *International Journal of Radiation Oncology*Biology*Physics*, vol. 105, no. 3, pp. 495–503, 2019. [Online]. Available: https://doi.org/10.1016/j.ijrobp.2019.06.2535

[32] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5967–5976. [Online]. Available: https://doi.org/10.1109/CVPR.2017.632

[33] M. E. Korsholm, L. W. Waring, and J. M. Edmund, "A criterion for the reliable use of mri-only radiotherapy," *Radiation Oncology*, vol. 9, no. 1, p. 16, 2014. [Online]. Available: https://doi.org/10.1186/1748-717X-9-16

[34] P. Klages, I. Benslimane, S. Riyahi, J. Jiang, M. Hunt, J. O. Deasy *et al.*, "Patch-based generative adversarial neural network models for head and neck MR-only planning," *Medical Physics*, vol. 47, no. 2, pp. 626–642, 2020. [Online]. Available: https://doi.org/10.1002/mp.13927

[35] L. M. O'Connor, J. H. Choi, J. A. Dowling, H. Warren-Forward, J. Martin, and P. B. Greer, "Comparison of synthetic computed tomography generation methods, incorporating male and female anatomical differences, for magnetic resonance imaging-only definitive pelvic radiotherapy," *Frontiers in Oncology*, vol. 12, no. 822687, 2022. [Online]. Available: https://doi.org/10.3389/fonc.2022.822687

[36] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian *et al.*, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017. [Online]. Available: https://doi.org/10.1016/j.media.2017.07.005

[37] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015. [Online]. Available: https://doi.org/10.1038/nature14539

[38] D. Nie, X. Cao, Y. Gao, L. Wang, and D. Shen, "Estimating CT Image from MRI Data Using 3D Fully Convolutional Networks," in *Deep Learning and Data Labeling for Medical Applications*, G. Carneiro, D. Mateus, L. Peter, A. Bradley, J. M. R. S. Tavares, V. Belagiannis *et al.*, Eds., 2016, pp. 170–178. [Online]. Available: https://doi.org/10.1007/978-3-319-46976-8_18

[39] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair *et al.*, "Generative Adversarial Nets," in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, ser. NIPS'14, 2014, pp. 2672–2680. [Online]. Available: https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf

[40] D. Nie, R. Trullo, J. Lian, C. Petitjean, S. Ruan, Q. Wang *et al.*, "Medical Image Synthesis with Context-Aware Generative Adversarial Networks," in *Medical Image Computing and Computer Assisted Intervention (MICCAI) 2017*, M. Descoteaux, L. Maier-Hein, A. Franz, P. Jannin, D. L. Collins, and S. Duchesne, Eds., 2017, pp. 417–425. [Online]. Available: https://doi.org/10.1007/978-3-319-66179-7_48

[41] J. M. Wolterink, A. M. Dinkla, M. H. F. Savenije, P. R. Seevinck, C. A. T. van den Berg, and I. Išgum, "Deep MR to CT synthesis using unpaired data," in *Simulation and Synthesis in Medical Imaging*, ser. Lecture Notes in Computer Science (LNCS), S. A. Tsaftaris, A. Gooya, A. F. Frangi, and J. L. Prince, Eds., vol. 10557, 2017, pp. 14–23. [Online]. Available: https://doi.org/10.1007/978-3-319-68127-6_2

[42] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2242–2251. [Online]. Available: https://doi.org/10.1109/ICCV.2017.244

[43] J. Fu, K. Singhrao, M. Cao, V. Yu, A. P. Santhanam, Y. Yang *et al.*, "Generation of abdominal synthetic CTs from 0.35T MR images using generative adversarial networks for MR-only liver radiotherapy," *Biomedical Physics and Engineering Express*, vol. 6, no. 1, p. 015033, 2020. [Online]. Available: https://doi.org/10.1088/2057-1976/ab6e1f

[44] L. Liu, A. Johansson, Y. Cao, J. Dow, T. S. Lawrence, and J. M. Balter, "Abdominal synthetic CT generation from MR Dixon images using a U-net trained with 'semi-synthetic' CT data," *Physics in Medicine and Biology*, vol. 65, no. 12, p. 125001, 2020. [Online]. Available: https://doi.org/10.1088/1361-6560/ab8cd2

[45] A. M. Dinkla, J. M. Wolterink, M. Maspero, M. H. F. Savenije, J. J. C. Verhoeff, E. Seravalli *et al.*, "MR-Only Brain Radiation Therapy: Dosimetric Evaluation of Synthetic CTs Generated by a Dilated Convolutional Neural Network," *International Journal of Radiation Oncology*Biology*Physics*, vol. 102, no. 4, pp. 801–812, 2018. [Online]. Available: https://doi.org/10.1016/j.ijrobp.2018.05.058

[46] S. Olberg, H. Zhang, W. R. Kennedy, J. Chun, V. Rodriguez, I. Zoberi *et al.*, "Synthetic CT reconstruction using a deep spatial pyramid convolutional framework for MR-only breast radiotherapy," *Medical Physics*, vol. 46, no. 9, pp. 4135–4147, 2019. [Online]. Available: https://doi.org/10.1002/mp.13716

[47] M. L. Groot Koerkamp, Y. J. M. De Hond, M. Maspero, C. Kontaxis, S. Mandija, J. E. Vasmel *et al.*, "Synthetic CT for single-fraction neoadjuvant partial breast irradiation on an MRI-linac," *Physics in Medicine and Biology*, vol. 66, no. 8, p. 085010, 2021. [Online]. Available: https://doi.org/10.1088/1361-6560/abf1ba

[48] M. Maspero, M. H. F. Savenije, A. M. Dinkla, P. R. Seevinck, M. P. W. Intven, I. M. Jurgenliemk-Schulz *et al.*, "Dose evaluation of fast synthetic-CT generation using a generative adversarial network for general pelvis MR-only radiotherapy," *Physics in Medicine and Biology*, vol. 63, no. 18, p. 185001, 2018. [Online]. Available: https://doi.org/10.1088/1361-6560/aada6d

[49] K. N. D. Brou Boni, J. Klein, A. Gulyban, N. Reynaert, and D. Pasquier, "Improving generalization in MR-to-CT synthesis in radiotherapy by using an augmented cycle generative adversarial network with unpaired data," *Medical Physics*, vol. 48, no. 6, pp. 3003–3010, 2021. [Online]. Available: https://doi.org/10.1002/mp.14866

[50] S. Chen, A. Qin, D. Zhou, and D. Yan, "Technical Note: U-net-generated synthetic CT images for magnetic resonance imaging-only prostate intensity-modulated radiation therapy treatment planning," *Medical Physics*, vol. 45, no. 12, pp. 5659–5665, 2018. [Online]. Available: https://doi.org/10.1002/mp.13247

[51] D. Bird, M. G. Nix, H. McCallum, M. Teo, A. Gilbert, N. Casanova *et al.*, "Multicentre, deep learning, synthetic-CT generation for ano-rectal MR-only radiotherapy treatment planning," *Radiotherapy and Oncology*, vol. 156, pp. 23–28, 2021. [Online]. Available: https://doi.org/10.1016/j.radonc.2020.11.027

[52] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010. [Online]. Available: https://doi.org/10.1109/TKDE.2009.191

[53] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar, "Do imagenet classifiers generalize to imagenet?" in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research (PMLR), K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97, 2019, pp. 5389–5400. [Online]. Available: https://proceedings.mlr.press/v97/recht19a.html

[54] J. G. Moreno-Torres, T. Raeder, R. Alaiz-Rodríguez, N. V. Chawla, and F. Herrera, "A unifying view on dataset shift in classification," *Pattern Recognition*, vol. 45, no. 1, pp. 521–530, 2012. [Online]. Available: https://doi.org/10.1016/j.patcog.2011.06.019

[55] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017, pp. 23–30. [Online]. Available: https://doi.org/10.1109/IROS.2017.8202133

[56] W. Li, S. Kazemifar, T. Bai, D. Nguyen, Y. Weng, Y. Li *et al.*, "Synthesizing CT images from MR images with deep learning: Model generalization for different datasets through transfer learning," *Biomedical Physics and Engineering Express*, vol. 7, no. 2, p. 025020, 2021. [Online]. Available: https://doi.org/10.1088/2057-1976/abe3a7

[57] L. Zimmermann, B. Knäusl, M. Stock, C. Lütgendorf-Caucig, D. Georg, and P. Kuess, "An mri sequence independent convolutional neural network for synthetic head ct generation in proton therapy," *Zeitschrift fur Medizinische Physik*, 2021. [Online]. Available: https://doi.org/10.1016/j.zemedi.2021.10.003

[58] B. Billot, D. N. Greve, K. Van Leemput, B. Fischl, J. E. Iglesias, and A. Dalca, "A Learning Strategy for Contrast-agnostic MRI Segmentation," in *Proceedings of the Third Conference on Medical Imaging with Deep Learning (MIDL)*, ser. Proceedings of Machine Learning Research (PMLR), T. Arbel, I. Ben Ayed, M. de Bruijne, M. Descoteaux, H. Lombaert, and C. Pal, Eds., vol. 121, 2020, pp. 75–93. [Online]. Available: https://proceedings.mlr.press/v121/billot20a.html

[59] B. Billot, D. N. Greve, O. Puonti, A. Thielscher, K. V. Leemput, B. Fischl *et al.*, "Synthseg: Domain randomisation for segmentation of brain scans of any contrast and resolution," 2021. [Online]. Available: http://arxiv.org/abs/2107.09559

[60] Y. Bengio, F. Bastien, A. Bergeron, N. Boulanger–Lewandowski, T. Breuel, Y. Chherawala *et al.*, "Deep learners benefit more from out-of-distribution examples," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research (PMLR), G. Gordon, D. Dunson, and M. Dudík, Eds., vol. 15, 2011, pp. 164–172. [Online]. Available: https://proceedings.mlr.press/v15/bengio11b.html

[61] M. Hoffmann, B. Billot, J. E. Iglesias, B. Fischl, and A. V. Dalca, "Learning mri contrast-agnostic registration," in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, vol. 2021-April, 2021, pp. 899–903. [Online]. Available: https://doi.org/10.1109/ISBI48211.2021.9434113

[62] J. E. Iglesias, B. Billot, Y. Balbastre, A. Tabari, J. Conklin, R. Gilberto González *et al.*, "Joint super-resolution and synthesis of 1 mm isotropic MP-RAGE volumes from clinical MRI exams with scans of different orientation, resolution and contrast," *NeuroImage*, vol. 237, p. 118206, 2021. [Online]. Available: https://doi.org/10.1016/j.neuroimage.2021.118206

[63] M. Maspero, L. G. Bentvelzen, M. H. F. Savenije, F. Guerreiro, E. Seravalli, G. O. Janssens *et al.*, "Deep learning-based synthetic CT generation for paediatric brain MR-only photon and proton radiotherapy," *Radiotherapy and Oncology*, vol. 153, pp. 197–204, 2020. [Online]. Available: https://doi.org/10.1016/j.radonc.2020.09.029

[64] K. N. D. Brou Boni, J. Klein, L. Vanquin, A. Wagner, T. Lacornerie, D. Pasquier *et al.*, "MR to CT synthesis with multicenter data in the pelvic area using a conditional generative adversarial network," *Physics in Medicine and Biology*, vol. 65, no. 7, p. 075002, 2020. [Online]. Available: https://doi.org/10.1088/1361-6560/ab7633

[65] A. Olin, C. Thomas, A. Hansen, J. Rasmussen, G. Krokos, T. Urbano *et al.*, "Robustness and generalizability of deep learning synthetic computed tomography for positron emission tomography/magnetic resonance imaging–based radiation therapy planning of patients with head and neck cancer," *Advances in Radiation Oncology*, vol. 6, 2021. [Online]. Available: https://doi.org/10.1016/j.adro.2021.100762

[66] L. Fetty, T. Löfstedt, G. Heilemann, H. Furtado, N. Nesvacil, T. Nyholm *et al.*, "Investigating conditional GAN performance with different generator architectures, an ensemble model, and different MR scanners for MR-sCT conversion," *Physics in Medicine and Biology*, vol. 65, no. 10, p. 105004, 2020. [Online]. Available: https://doi.org/10.1088/1361-6560/ab857b

[67] C. Wang, J. Uh, X. He, C.-H. Hua, and S. Acharya, "Transfer learning-based synthetic ct generation for mr-only proton therapy planning in children with pelvic sarcomas," in *Medical Imaging 2021: Physics of Medical Imaging*, ser. Proceedings of SPIE, vol. 11595, 2021. [Online]. Available: https://doi.org/10.1117/12.2579767

[68] C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," *Journal of Big Data*, vol. 6, no. 1, p. 60, 2019. [Online]. Available: https://doi.org/10.1186/s40537-019-0197-0

[69] A. Sanaat, I. Shiri, S. Ferdowsi, H. Arabi, and H. Zaidi, "Robust-Deep: A Method for Increasing Brain Imaging Datasets to Improve Deep Learning Models' Performance and Robustness," *Journal of Digital Imaging*, 2022. [Online]. Available: https://doi.org/10.1007/s10278-021-00536-0

[70] A. Almahairi, S. Rajeswar, A. Sordoni, P. Bachman, and A. Courville, "Augmented cyclegan: Learning many-to-many mappings from unpaired data," 2018. [Online]. Available: https://arxiv.org/abs/1802.10151

[71] M. Gadermayr, M. Tschuchnig, L. Gupta, N. Kramer, D. Truhn, D. Merhof *et al.*, "An asymmetric cycle-consistency loss for dealing with many-to-one mappings in image translation: A study on thigh mr scans," in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, 2021, pp. 1182–1186. [Online]. Available: https://doi.org/10.1109/ISBI48211.2021.9433891

[72] S. Klein, M. Staring, K. Murphy, M. A. Viergever, and J. P. W. Pluim, "elastix: A Toolbox for Intensity-Based Medical Image Registration," *IEEE Transactions on Medical Imaging*, vol. 29, no. 1, pp. 196–205, 2010. [Online]. Available: https://doi.org/10.1109/TMI.2009.2035616

[73] D. Shamonin, E. Bron, B. Lelieveldt, M. Smits, S. Klein, and M. Staring, "Fast Parallel Image Registration on CPU and GPU for Diagnostic Classification of Alzheimer's Disease," *Frontiers in Neuroinformatics*, vol. 7, p. 50, 2014. [Online]. Available: https://doi.org/10.3389/fninf.2013.00050

[74] S. Hissoiny, B. Ozell, H. Bouchard, and P. Després, "GPUMCD: A new GPU-oriented Monte Carlo dose calculation platform," *Medical Physics*, vol. 38, pp. 754–764, 2011, https://doi.org/10.1118/1.3539725. [Online]. Available: https://doi.org/10.1118/1.3539725

[75] D. A. Low, W. B. Harms, S. Mutic, and J. Purdy, "A technique for the quantitative evaluation of dose distributions." *Medical Physics*, vol. 25, no. 5, pp. 656–61, 1998. [Online]. Available: https://doi.org/10.1118/1.598248

[76] G. Heilemann, B. Poppe, and W. U. Laub, "On the sensitivity of common gamma-index evaluation methods to mlc misalignments in rapidarc quality assurance." *Medical physics*, vol. 40, no. 3, p. 031702, 2013. [Online]. Available: https://doi.org/10.1118/1.4789580

[77] I. Hadzic, S. Pai, R. Chinmay, and J. Teuwen, "ganslate-team/ganslate: Initial public release," 2021. [Online]. Available: https://doi.org/10.5281/zenodo.5494572

[78] D. Kingma and J. Ba, "ADAM: A Method for Stochastic Optimization," in *Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015)*, 2015, pp. 1–15. [Online]. Available: https://arxiv.org/abs/1412.6980

[79] L. Henschel, S. Conjeti, S. Estrada, K. Diers, B. Fischl, and M. Reuter, "Fastsurfer - a fast and accurate deep learning based neuroimaging pipeline," *NeuroImage*, vol. 219, p. 117012, 2020. [Online]. Available: https://doi.org/10.1016/j.neuroimage.2020.117012

[80] M. H. F. Savenije, M. Maspero, G. G. Sikkes, J. R. N. van der Voort van Zyp, A. N. T. J. Kotte, G. H. Bol *et al.*, "Clinical implementation of MRI-based organs-at-risk auto-segmentation with convolutional networks for prostate radiotherapy," *Radiation Oncology*, vol. 15, no. 1, p. 104, 2020. [Online]. Available: https://doi.org/10.1186/s13014-020-01528-0

[81] K. Kamnitsas, C. Ledig, V. F. J. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon *et al.*, "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation," *Medical Image Analysis*, vol. 36, pp. 61–78, 2017. [Online]. Available: https://doi.org/10.1016/j.media.2016.10.004

[82] F. Pérez-García, R. Sparks, and S. Ourselin, "TorchIO: A Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning," *Computer Methods and Programs in Biomedicine*, vol. 208, p. 106236, 2021. [Online]. Available: https://doi.org/10.1016/j.cmpb.2021.106236

[83] H. Massa, J. Johnson, and A. McMillan, "Comparison of deep learning synthesis of synthetic CTs using clinical MRI inputs," *Physics in Medicine and Biology*, vol. 65, 2020. [Online]. Available: https://doi.org/10.1088/1361-6560/abc5cb

[84] S. Irmak, L. Zimmermann, D. Georg, P. Kuess, and W. Lechner, "Cone beam CT based validation of neural network generated synthetic CTs for radiotherapy in the head region," *Medical Physics*, vol. 48, no. 8, pp. 4560–4571, 2021. [Online]. Available: https://doi.org/10.1002/mp.14987

[85] H. Emami, M. Dong, S. Nejad-Davarani, and C. Glide-Hurst, "Generating Synthetic CTs from Magnetic Resonance Images using Generative Adversarial Networks," *Medical Physics*, vol. 45, 6 2018. [Online]. Available: https://doi.org/10.1002/mp.13047

[86] E. A. Andres, L. Fidon, M. Vakalopoulou, M. Lerousseau, A. Carré, R. Sun *et al.*, "Dosimetry-Driven Quality Measure of Brain Pseudo Computed Tomography Generated From Deep Learning for MRI-Only Radiation Therapy Treatment Planning," *International Journal of Radiation Oncology Biology Physics*, vol. 108, pp. 813–823, 2020. [Online]. Available: https://doi.org/10.1016/j.ijrobp.2020.05.006

[87] C. Wang, J. Uh, T. Patni, T. Merchant, Y. Li, C.-H. Hua *et al.*, "Toward mr-only proton therapy planning for pediatric brain tumors: Synthesis of relative proton stopping power images with multiple sequence mri and development of an online quality assurance tool," *Medical Physics*, vol. 49, no. 3, pp. 1559–1570, 2022. [Online]. Available: https://doi.org/10.1002/mp.15479

[88] M. C. Florkow, F. Zijlstra, L. G. W. Kerkmeijer, M. Maspero, C. A. T. Van Den Berg, M. Van Stralen *et al.*, "The impact of MRI-CT registration errors on deep learning-based synthetic CT generation," in *Medical Imaging 2019: Image Processing*, ser. Proceedings of SPIE, vol. 10949, 2019. [Online]. Available: https://doi.org/10.1117/12.2512747

[89] J. Tremblay, A. Prakash, D. Acuna, M. Brophy, V. Jampani, C. Anil *et al.*, "Training deep networks with synthetic data: Bridging the reality gap by domain randomization," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018, pp. 1082–10 828.

[90] F. C. Nogues, A. Huie, and S. Dasgupta, "Object detection using domain randomization and generative adversarial refinement of synthetic images," 2018. [Online]. Available: https://arxiv.org/abs/1805.11778

[91] V. Sandfort, K. Yan, P. J. Pickhardt, and R. M. Summers, "Data augmentation using generative adversarial networks (cyclegan) to improve generalizability in ct segmentation tasks," *Scientific Reports*, vol. 9, p. 16884, 2019. [Online]. Available: https://doi.org/10.1038/s41598-019-52737-x

[92] Y. Liu, Y. Lei, Y. Wang, G. Shafai-Erfani, T. Wang, S. Tian *et al.*, "Evaluation of a deep learning-based pelvic synthetic CT generation technique for MRI-based prostate proton treatment planning," *Physics in Medicine and Biology*, vol. 64, no. 20, p. 205022, 2019. [Online]. Available: https://doi.org/10.1088/1361-6560/ab41af

[93] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004. [Online]. Available: https://doi.org/10.1109/TIP.2003.819861

[94] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in *International Conference on Learning Representations (ICLR)*, 2019. [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7

[95] S. Neppl, G. Landry, C. Kurz, D. C. Hansen, B. Hoyle, S. Stöcklein *et al.*, "Evaluation of proton and photon dose distributions recalculated

on 2D and 3D Unet-generated pseudoCTs from T1-weighted MR head scans," *Acta Oncologica*, vol. 58, no. 10, pp. 1429–1434, 2019. [Online]. Available: https://doi.org/10.1080/0284186X.2019.1630754

[96] J. Fu, Y. Yang, K. Singhrao, D. Ruan, F.-I. Chu, D. A. Low *et al.*, "Deep learning approaches using 2D and 3D convolutional neural networks for generating male pelvic synthetic computed tomography from magnetic resonance imaging," *Medical Physics*, vol. 46, no. 9, pp. 3788–3798, 2019. [Online]. Available: https://doi.org/10.1002/mp.13672

[97] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979. [Online]. Available: https://doi.org/10.1109/TSMC.1979.4310076

[98] L. Xu, X. Zeng, H. Zhang, W. Li, J. Lei, and Z. Huang, "BPGAN: Bidirectional CT-to-MRI prediction using multi-generative multi-adversarial nets with spectral normalization and localization," *Neural Networks*, vol. 128, pp. 82–96, 2020. [Online]. Available: https://doi.org/10.1016/j.neunet.2020.05.001

[99] A. Jabbarpour, S. Mahdavi, A. V. Sadr, G. Esmaili, I. Shiri, and H. Zaidi, "Unsupervised pseudo CT generation using heterogenous multicentric CT/MR images and CycleGAN: Dosimetric assessment for 3D conformal radiotherapy," *Computers in Biology and Medicine*, vol. 143, 2022. [Online]. Available: https://doi.org/10.1016/j.compbiomed.2022.105277

[100] X. Han, "MR-based synthetic CT generation using a deep convolutional neural network method," *Medical Physics*, vol. 44, no. 4, pp. 1408–1419, 2017. [Online]. Available: https://doi.org/10.1002/mp.12155

[101] L. Xiang, Q. Wang, D. Nie, L. Zhang, X. Jin, Y. Qiao *et al.*, "Deep embedding convolutional neural network for synthesizing CT image from T1-Weighted MR image," *Medical Image Analysis*, vol. 47, pp. 31–44, 2018. [Online]. Available: https://doi.org/10.1016/j.media.2018.03.011

[102] D. Gupta, M. Kim, K. Vineberg, and J. Balter, "Generation of synthetic CT images from MRI for treatment planning and patient positioning using a 3-channel U-net trained on sagittal images," *Frontiers in Oncology*, vol. 9, 2019. [Online]. Available: https://doi.org/10.3389/fonc.2019.00964

[103] Y. Koike, Y. Akino, I. Sumida, H. Shiomi, H. Mizuno, M. Yagi *et al.*, "Feasibility of synthetic computed tomography generated with an adversarial network for multi-sequence magnetic resonance-based brain radiotherapy," *Journal of Radiation Research*, vol. 61, pp. 92–103, 2019. [Online]. Available: https://doi.org/10.1093/jrr/rrz063

[104] Y. Lei, J. Harms, T. Wang, Y. Liu, H.-K. Shu, A. Jani *et al.*, "MRI-Only Based Synthetic CT Generation Using Dense Cycle Consistent Generative Adversarial Networks," *Medical Physics*, 2019. [Online]. Available: https://doi.org/10.1002/mp.13617

[105] G. Shafai-Erfani, Y. Lei, Y. Liu, Y. Wang, T. Wang, J. Zhong *et al.*, "MRI-Based Proton Treatment Planning for Base of Skull Tumors," *International Journal of Particle Therapy*, vol. 6, pp. 12–25, 2019. [Online]. Available: https://doi.org/10.14338/IJPT-19-00062.1

[106] S. Sreeja and D. Mubarak, "Pseudo Computed Tomography Image Generation from Brain Magnetic Resonance Image for Radiation Therapy Treatment Planning Using DCNN-UNET," *Webology*, vol. 18, pp. 704–726, 2021. [Online]. Available: https://doi.org/10.14704/WEB/V18SI05/WEB18256

[107] B. Tang, F. Wu, Y. Fu, X. Wang, P. Wang, L. Orlandini *et al.*, "Dosimetric evaluation of synthetic ct image generated using a neural network for mr-only brain radiotherapy," *Journal of Applied Clinical Medical Physics*, vol. 22, pp. 55–62, 2021. [Online]. Available: https://doi.org/10.1002/acm2.13176

[108] F. Gholamiankhah, S. Mostafapour, and H. Arabi, "Deep learning-based synthetic ct generation from mr images: comparison of generative adversarial and residual neural networks," *International Journal of Radiation Research*, vol. 20, no. 1, pp. 121–130, 2022. [Online]. Available: https://doi.org/10.52547/ijrr.20.1.19

[109] X. Liu, H. Emami, S. Nejad-Davarani, E. Morris, L. Schultz, M. Dong *et al.*, "Performance of deep learning synthetic CTs for MR-only brain radiation therapy," *Journal of Applied Clinical Medical Physics*, vol. 22, no. 1, pp. 308–317, 2021. [Online]. Available: https://doi.org/10.1002/acm2.13139

[110] S. Kazemifar, S. McGuire, R. Timmerman, Z. Wardak, D. Nguyen, Y. Park *et al.*, "MRI-only brain radiotherapy: Assessing the dosimetric accuracy of synthetic CT images generated using a deep learning approach," *Radiotherapy and Oncology*, vol. 136, pp. 56–63, 2019. [Online]. Available: https://doi.org/10.1016/j.radonc.2019.03.026

[111] F. Liu, P. Yadav, A. Baschnagel, and A. McMillan, "MR-based treatment planning in radiation therapy using a deep learning approach," *Journal of Applied Clinical Medical Physics*, vol. 20, no. 3, pp. 105–114, 2019. [Online]. Available: https://doi.org/10.1002/acm2.12554

[112] Y. Li, W. Li, J. Xiong, J. Xia, and Y. Xie, "Comparison of Supervised and Unsupervised Deep Learning Methods for Medical Image Synthesis between Computed Tomography and Magnetic Resonance Images," *BioMed Research International*, vol. 2020, p. 5193707, 2020. [Online]. Available: https://doi.org/10.1155/2020/5193707

[113] A. Ranjan, D. Lalwani, and R. Misra, "GAN for synthesizing CT from T2-weighted MRI data towards MR-guided radiation treatment," *Magnetic Resonance Materials in Physics, Biology and Medicine*, 2021. [Online]. Available: https://doi.org/10.1007/s10334-021-00974-5

APPENDIX

A. IMAGE SIMILARITY METRICS

This section gives the definitions of the image similarity metrics used in this work. MAE is defined as [27]:

$$MAE = \frac{\sum_1^n |CT_i - sCT_i|}{n}, \tag{6}$$

with n the number of voxels in the given region of interest (ROI). The metric gives an impression of the overall discrepancy in the assigned HU value between the acquired CT image and the sCT [92]. Lower values mean better congruence between sCT and acquired CT.

PSNR is defined as follows [27]:

$$PSNR = 10 * \log_{10} \frac{MAX_{CT}^2}{MSE}, \tag{7}$$

in which MAX_{CT} is the maximum possible intensity value within the CT. Here, MSE is the mean square error: $MSE = \frac{\sum_1^n (CT_i - sCT_i)^2}{n}$. PSNR quantifies the noise introduced by the synthesis of CT compared to using the acquired CT image [27]. Better performance of a model used for sCT generation translates to a higher PSNR.

SSIM is computed to add an element of perception. The metric considers known attributes of the human visual system, where quality is defined in terms of decay in image structure [93]. A higher value means the generated sCT is more similar to the acquired CT image. The metric is computed as [93]:

$$SSIM = \frac{(2\mu_{sCT}\mu_{CT} + c_1)(2\sigma_{sCT,CT} + c_2)}{(\mu_{sCT}^2 + \mu_{CT}^2 + c_1)((\mu_{sCT}^2 + \sigma_{CT}^2 + c_2)}, \tag{8}$$

with $c_1 = (k_1 L)^2$ and $c_2 = (k_2 L)^2$. L is the dynamic range in the image, μ is the mean value, σ is the (co)variance and $k_{1,2}$ are constants: $k_1 = 0.01$ and $k_2 = 0.03$.

B. EXAMPLE IMAGES OF WATER-FILLING FOR RT PLAN RE-CALCULATION

The difference in FOV of the acquired MRI and corresponding CT led to equal differences in FOV between sCT and acquired CT. This difference was water-filled in both sCT and acquired CT for the dose re-calculation performed for assessment of dosimetric accuracy of sCT-based dose plans by setting the voxels outside the FOV of the original MRI but inside the body contour of the original acquired CT equal to 0 HU, as illustrated in Fig. 15.

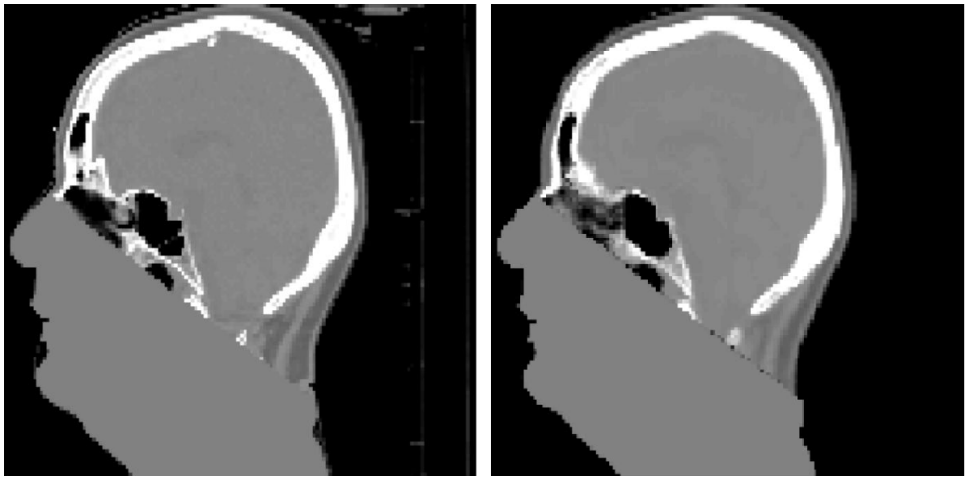

Fig. 15. The CT_{wf} (left) and sCT_{wf} (right) generated by the Baseline model from the T1w image of an example patient in the test set after water-filling the original images for RT plan re-calculation.

C. DOSIMETRIC ACCURACY: GAMMA INDEX AND PASS RATE

This section describes how the γ-index and -pass rate can be calculated. The γ-pass rate combines measures for spatial distance and dose difference. In the computation of the γ-index, a comparison is made between a reference dose distribution (here: the dose distribution obtained through planning with acquired CT) and the dose distribution that needs to be evaluated (here: the dose distribution resulting from planning based on sCT).

For each point r_e in the distribution for evaluation, the γ-index is computed as follows [75]:

$$\gamma(r_e, r_r) = \sqrt{\frac{\Delta r^2(r_e, r_r)}{\delta r} + \frac{\Delta D^2(r_e, r_r)}{\delta D}}. \tag{9}$$

Here, r_r is a given point in the reference distribution, Δr is the Euclidean distance between r_e and r_r in space, and ΔD is the difference between the dose in each of the two points (eq. 10).

$$\Delta D(r_e, r_r) = D_e(r_e) - D_r(r_r) \tag{10}$$

For each point r_r, a value for γ is obtained by minimising over all points r_e.

In eq. 9, δD and δr are the dose and distance criteria, respectively, usually reported as, e.g., δD/δr = 3%/3mm. A point in the reference distribution passes this criterion if $\gamma \leq 1$, so a γ-pass rate can be computed.

D. Model optimisation: hyperparameter tuning

Hyperparameter optimisation was introduced in section III-E. Details about the hyperparameter optimisation process are provided here, including a summary of the main results.

A. Methods

Hyperparameter optimisation was performed for a 2D and a 3D model in parallel. Tuning was done using a subset of 10 patients from the training dataset and the validation set (n = 10), using only T1w images without Gadolinium contrast. The optimisation was done by training 50 epochs for the 2D configuration or 5,000 (coarse search) or 25,000 (refined search) iterations for the 3D configuration. Here, an epoch is defined as passing the entire training set through the network once. An iteration is defined as passing one image patch per patient through the network times the number of patients in one batch. The MAE between CT_{crop} and generated sCT was leading in optimisation. SSIM and PSNR were used to decide if no differences in MAE were observed between models. Additionally, images were visually inspected for artefacts and image quality.

Hyperparameter optimisation was done through a grid search strategy. Hyperparameters considered were: (1) *initialisation method*, (2) *optimiser* and (3) corresponding *weight decay* value for the AdamW optimiser [94], (4) *patch size* for the 3D configuration or *load size* for the 2D configuration, (5) the value of λ in the loss function, (6) *learning rate*, (7) *batch size*, and (8) *number of downsampling steps* used in the U-Net generator architecture. Specifically for the 3D configuration, where patch-based inference was used, (9) the *blend mode* for patch combination and (10) the amount of *patch overlap* were tuned. Table V contains the grid values considered.

TABLE V
HYPERPARAMETERS CONSIDERED FOR OPTIMISATION AND THE CORRESPONDING TESTED GRID VALUES.

Hyperparameter	**2D configuration** (n)	**3D configuration** (n)
Initialisation method	[Kaiming, Xavier]	[Kaiming, Xavier]
Optimiser	[Adam, AdamW]	[Adam, AdamW]
Weight decay for AdamW	[0.01, 0.1, 0.5]	[0.01, 0.1, 0.5]
Load size (2D), Patch size (3D)	[256, 268, 286, 512, 1024]	[32, 64, 128, 256]
λ	[100, 500, 1,000, 5,000, 10,000]	[1, 10, 100, 500, 1,000, 5,000, 10,000]
Learning rate	[0.0001, 0.001, 0.005]	[0.0001, 0.001, 0.01]
Batch size	[1, 2, 5, 10, 50]	[1, 5, 10]
Number of downsampling steps	[5, 6, 7, 8]	[5, 6, 7]
Blend mode	-	[Gaussian, constant]
Patch overlap	-	[0.25, 0.5, 0.75]

As a final optimisation step, three different decay strategies for the learning rate were compared for both configurations, and early stopping was investigated. In this step, training was done on 30 training patients. The decay strategies considered were a constant learning rate of 0.001, a stepwise decaying learning rate and a cyclic learning rate. Two decay steps were applied for the stepwise decaying learning rate with decay factor γ = 0.2, after 40 and 60 epochs for the 2D model and 200,000 and 300,000 iterations for the 3D model. The cyclic learning rate was implemented with an initial constant phase at a learning rate of 0.001 during 40 epochs (2D) or 200,000 iterations (3D), followed by linear decay to 0 in 40 epochs (2D) or 200,000 iterations (3D) and two cycles consisting of a restart to a learning rate of 0.01 with decay to 0 in 60 epochs (2D) or 300,000 iterations (3D). Early stopping was applied by selecting the first epoch or iteration for which the MAE in the intersection of the body contours did not improve for the subsequent three epochs or iterations. Note that evaluation was performed every 20 epochs for 2D or 20,000 iterations for 3D.

B. Results

This subsection summarises the main results of hyperparameter optimisation for the 2D and 3D models. The results are presented as a plot of SSIM against MAE for a selection of the models trained during several hyperparameter optimisation steps. Results are provided for the hyperparameters that proved to be the most influential: the learning rate, the value of λ and the patch size for 3D models or the load size for 2D models. All models for which the results are presented here were trained with Xavier initialisation, Adam optimiser (with momentum parameters β_1 = 0.5 and β_2 = 0.999 and no weight decay), batch size = 1, and a U-net generator architecture with 5 (3D models) or 8 (2D models) downsampling steps. Unless stated otherwise, the models for which the results for a given hyperparameter (λ, learning rate or patch or load size) are presented are those trained with the other hyperparameter values equal to those used in the final, optimised models. Models for which results are plotted in the left upper quadrant have the best performance: a lower MAE and a higher SSIM.

Figure 16 shows the SSIM plotted against the MAE for three models with a 3D configuration, trained with different learning rates. All models were trained during a first coarse optimisation step, training for 5,000 iterations with λ = 1,000 and a patch size = 128. Performance was best for the model with a learning rate = 0.001, and this value was adopted in the final model.

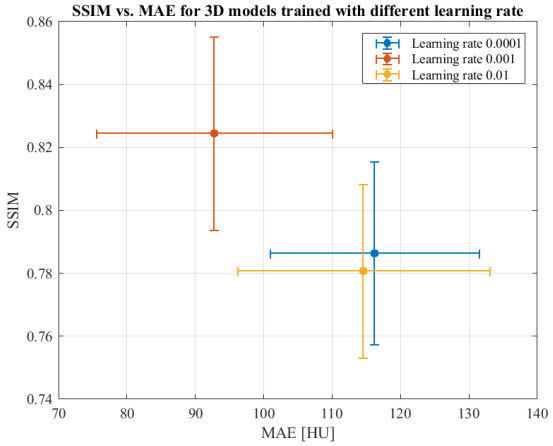

Fig. 16. Plot of the SSIM against the MAE [HU] for three models with 3D configuration trained with different learning rates. The marker represents the mean value of each metric. The error bars represent the standard deviation. All models were trained during a coarse initial optimisation step for 5,000 iterations with $\boldsymbol{\lambda}$ = 1,000 and a patch size = 128 on a subset of the training data constituting ten patients. Results were computed on the validation set (n = 10). Models with better performance are plotted toward the left upper corner: these have a high SSIM and a low MAE.

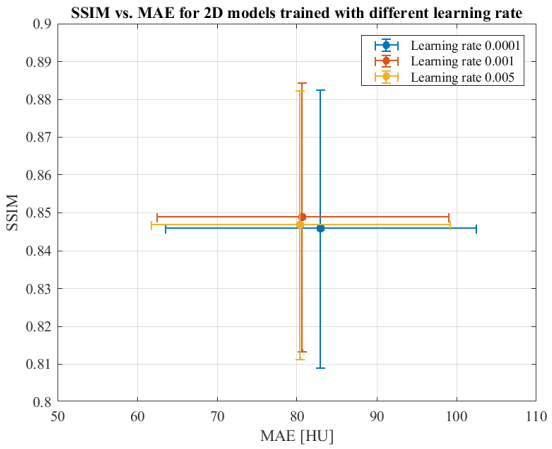

Fig. 17. Plot of the SSIM against the MAE [HU] for three models with 2D configuration trained with different learning rates. The marker represents the mean value of each metric. The error bars represent the standard deviation. All models were trained for 50 epochs with $\boldsymbol{\lambda}$ = 5,000 and a load size = 268 on a subset of the training data constituting ten patients. Results were computed on the validation set (n = 10). Models with better performance are plotted toward the left upper corner: these have a high SSIM and a low MAE.

For 2D models (Fig. 17), a learning rate = 0.001 resulted in better performance than a learning rate = 0.0001. A learning rate of 0.005 led to a marginally lower mean MAE than a learning rate of 0.001. However, the SSIM and PSNR (results for PSNR not shown) were better (higher) for the learning rate of 0.001, and the standard deviation in MAE was smaller for learning rate = 0.001. Additionally, other models in the grid search with different values for λ and the batch size showed a trend favouring a learning rate of 0.001 (results not shown), also when looking at mean MAE. Therefore, a learning rate of 0.001 was adopted for the 2D configuration. The models presented in Fig. 17 were all trained for 50 epochs with λ = 5,000 and load size = 268.

Figure 18 presents results for 3D models trained with different values of λ. Models indicated with a circular marker were trained during a first coarse optimisation step, training for 5,000 iterations with a learning rate = 0.001 and patch size = 128. The figure shows that increasing λ improved performance. Based on these results, the tested values of λ were broadened in a second grid search to include also λ = 5,000 and λ = 10,000. These models indicated with a triangular marker were trained for 25,000 iterations, with a learning rate = 0.001 and patch size = 128. Increasing λ improved performance up to λ = 5,000. An increase to λ = 10,000 did not improve performance further. During optimisation of λ for models with a 2D configuration (Fig. 19), the optimum was λ = 5,000 as well. Following these results, λ = 5,000 was chosen for both configurations.

Figure 20 presents results for 3D models trained with different patch sizes. Models indicated with a square marker were trained during a first coarse optimisation step, training for 5,000 iterations with a learning rate of 0.001 and λ = 1,000. A larger patch size resulted in improved performance. In a subsequent refined grid search, models were trained for 25,000 iterations with patch sizes of 128 and 256. Again, these models were trained with a learning rate of 0.001. For comparison, results are shown for models trained with λ = 1,000 and models trained with λ = 5,000. Increasing the patch size to 256 did not further improve performance. Therefore, a patch size of 128 was deemed optimal and was chosen. Figure 21 shows results for five 2D models trained with different load sizes. Images were subsequently cropped to a final size of 256 for all models, following the image size required for the U-Net-256 generator architecture. Load sizes 268 and 286 resulted in similar performances, with load size 268 slightly outperforming load size 286. Using the whole image (load size 256) for training or smaller patches (load sizes 512 or 1024) resulted in worse performance. A load size of 268 was chosen based on these results.

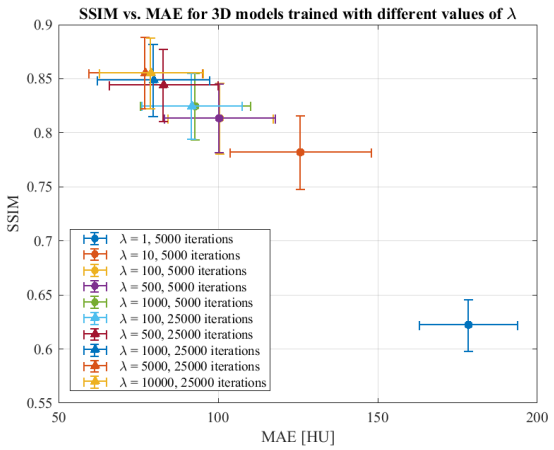

Fig. 18. Plot of the SSIM against the MAE [HU] for ten models with 3D configuration trained with different values of $\boldsymbol{\lambda}$. The marker represents the mean value of each metric. The error bars represent the standard deviation. All models were trained with a learning rate = 0.001 and patch size = 128 on a subset of the training data constituting ten patients. Models indicated with a circular marker were trained during an initial coarse optimisation step, training for 5,000 iterations. Models plotted with a triangular marker were trained during an optimisation step on a refined grid, training for 25,000 iterations. Results were computed on the validation set (n = 10). Models with better performance are plotted toward the left upper corner: these have a high SSIM and a low MAE.

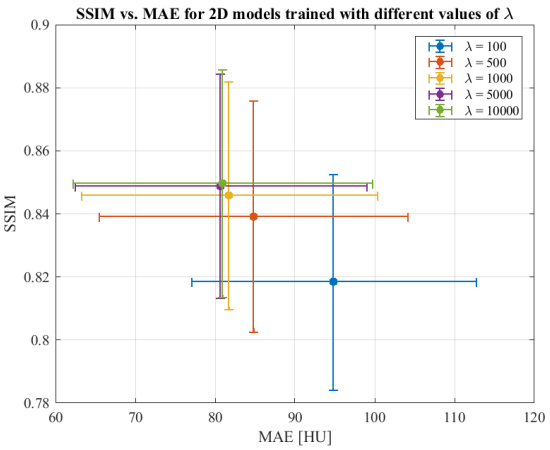

Fig. 19. Plot of the SSIM against the MAE [HU] for five models with 2D configuration trained with different values of $\boldsymbol{\lambda}$. The marker represents the mean value of each metric. The error bars represent the standard deviation. All models were trained for 50 epochs with a learning rate = 0.001 and load size = 268 on a subset of the training data comprising ten patients. Results were computed on the validation set (n = 10). Models with better performance are plotted toward the left upper corner: these have a high SSIM and a low MAE.

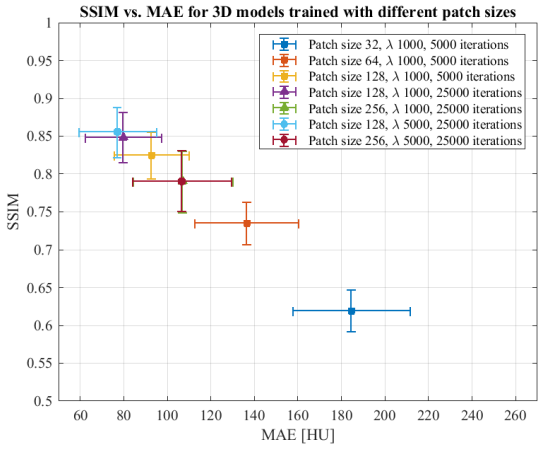

Fig. 20. Plot of the SSIM against the MAE [HU] for ten models with 3D configuration trained with different patch sizes. The marker represents the mean value of each metric. The error bars represent the standard deviation. All models were trained with a learning rate = 0.001 on a subset of the training data constituting ten patients. Models indicated with a square marker were trained during an initial coarse optimisation step, training for 5,000 iterations with $\boldsymbol{\lambda}$ = 1,000. Models plotted with a triangular marker were trained during an optimisation step on a refined grid, training for 25,000 iterations with $\boldsymbol{\lambda}$ = 1,000. Models plotted with a circular marker were trained during an optimisation step on a refined grid, training for 25,000 iterations with $\boldsymbol{\lambda}$ = 5,000. Results were computed on the validation set (n = 10). Models with better performance are plotted toward the left upper corner: these have a high SSIM and a low MAE.

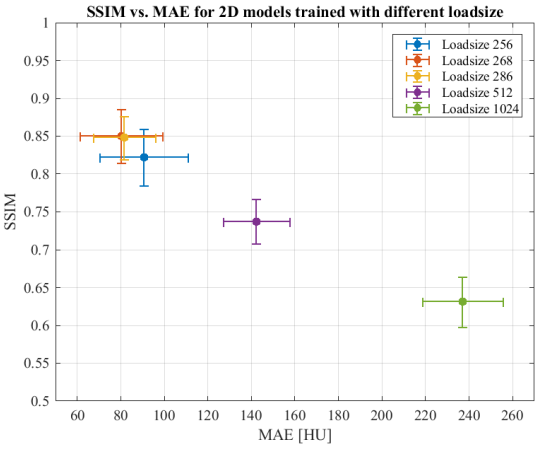

Fig. 21. Plot of the SSIM against the MAE [HU] for five models with 2D configuration trained with different load sizes. The marker represents the mean value of each metric. The error bars represent the standard deviation. All models were trained for 100 epochs with a learning rate = 0.001 and $\boldsymbol{\lambda}$ = 5,000 on a subset of the training data comprising ten patients. Results were computed on the validation set (n = 10). Models with better performance are plotted toward the left upper corner: these have a high SSIM and a low MAE.

E. Model optimisation: 2D vs 3D model configuration

As introduced in section III-E, the optimised 2D and 3D models were trained on a subset of thirty patients from the training set using T1w images only. Their performance was compared to establish a model configuration for all subsequent experiments. This section describes the methods and results for this comparison.

A. Methods

The optimised 2D model was trained using the following hyperparameters: Xavier initialisation, Adam optimiser, load size = 268x268 pixels (randomly cropped to a patch size of 256x256 pixels), batch size = 1, weight factor in the loss function λ = 5000, and 8 downsampling steps in the U-Net generator. A learning rate with stepwise decay was implemented, starting at an initial value of 0.001 and decaying with decay factor γ = 0.2 after 40 and 60 epochs. Early stopping was applied at epoch 60: the first epoch for which the MAE in the intersection of the body contours did not improve for the following three epochs, evaluating every 20 epochs.

The optimised 3D model was trained with Xavier initialisation, Adam optimiser, patch size = 128x128x128 voxels, batch size = 1, λ = 5000, number of downsampling steps = 5, and a constant learning rate of 0.001. A sliding window inferrer was used for patch combination with a patch overlap of 0.5 and Gaussian blend mode. Early stopping was applied at iteration 260,000. The same criterion was used as for the 2D model, evaluating every 20,000 iterations. The Adam optimiser [78] was used for both models with β_1 = 0.5 and β_2 = 0.999 as momentum parameters and no weight decay.

The performance of the optimised 2D and 3D models was compared by evaluating image similarity only on the T1w images of patients in the validation set (n = 10). Wilcoxon-signed rank tests were used to compare image similarity metrics between the two models statistically.

B. Results

The training times were 5.9 h and 31.3 h for the optimised 2D and 3D model, respectively. The 3D model significantly outperformed the 2D model for all three image similarity metrics (Table VI). Therefore, the 3D configuration was adopted.

TABLE VI
IMAGE SIMILARITY METRICS FOR SCT GENERATED BY THE OPTIMISED 2D AND 3D MODEL, COMPARED TO GROUND TRUTH CT.

Metric	*Model*		**p-value**
	2D	**3D**	
MAE [HU]	80.7 ± 14.3 [61.1 - 111]	71.3 ± 16.6 [55.3 - 111]	0.002
SSIM	0.848 ± 0.028 [0.794 - 0.882]	0.869 ± 0.033 [0.797 - 0.898]	0.002
PSNR [dB]	27.1 ± 1.45 [24.6 - 29.5]	28.0 ± 1.81 [24.6 - 30.4]	0.004

Metrics were calculated on T1w images of the validation set (n = 10) within the intersection of the body contour of the sCT and CT. Mean values and standard deviations ($\mu \pm 1\sigma$) and range ([min - max]) are reported. Wilcoxon-signed rank tests were used for statistical comparisons. Values of $p < 0.05$ were regarded as statistically significant.

C. Discussion

Few other studies have compared 2D and 3D networks for sCT generation for MR-only RT, obtaining mixed results. In [95], a 3D configuration decreased discontinuities between slices compared to a 2D configuration but increased blurriness in the output sCT images. Image similarity metrics and dosimetric comparisons favoured the 2D model [95]. In contrast, Fu *et al.* [96] reported results favouring a 3D model, in line with the findings in this work. Similar to [95], visual inspection of the network output revealed decreased discontinuities between slices for the 3D model compared to the 2D model in this work.

F. Model optimisation: hyperparameter finetuning and dataset balancing

This section explains the final optimisation step. The ratio between T1w images with/without contrast and T2w images in the training set was balanced, and the batch size was finetuned for the mix of input sequences. The ratios T1w:T1wGd:T2w = 1:1:2 or T1w:T1wGd:T2w = 1:1:1 were compared by re-training the 3D model on a subset of fifteen patients to balance the ratio between T1w images with/without contrast and T2w images in the training set. Values of 1 and 5 were considered for the batch size. In this step, decisions were based on the MAE obtained for the validation set, using T1w, T1wGd and T2w images, i.e., the seen sequences[3].

A batch size of 1 was chosen. The ratio of T1w:T1wGd:T2w = 1:1:2 was adopted for the training dataset, meaning the whole training dataset (n = 60 patients) contained 60 T2w images, 30 T1w images and 30 T1wGd images.

[3]One patient was retrospectively excluded from the validation set after failure of registration between the T2w image and the CT was observed. Validation of all models except the optimised 2D and 3D models was thus done on a nine-patient validation set.

G. Illustration of the early stopping method

This section illustrates how early stopping was applied throughout this work, using the early stopping of the Baseline model as a reference (Fig. 22). The MAE was calculated for T1w, T1wGd and T2w images in the intersection of the body contours. Additionally, a combined MAE was calculated as the average for the three sequences. The first iteration for which this combined MAE did not improve for the following three iterations was selected, evaluating every 50,000 iterations. In this case, early stopping was applied at iteration 300,000 (dashed lines).

For models trained using different acquired MRI (e.g., T1w(Gd) images only), the sequences taken into account when calculating the combined MAE were adjusted according to the training data.

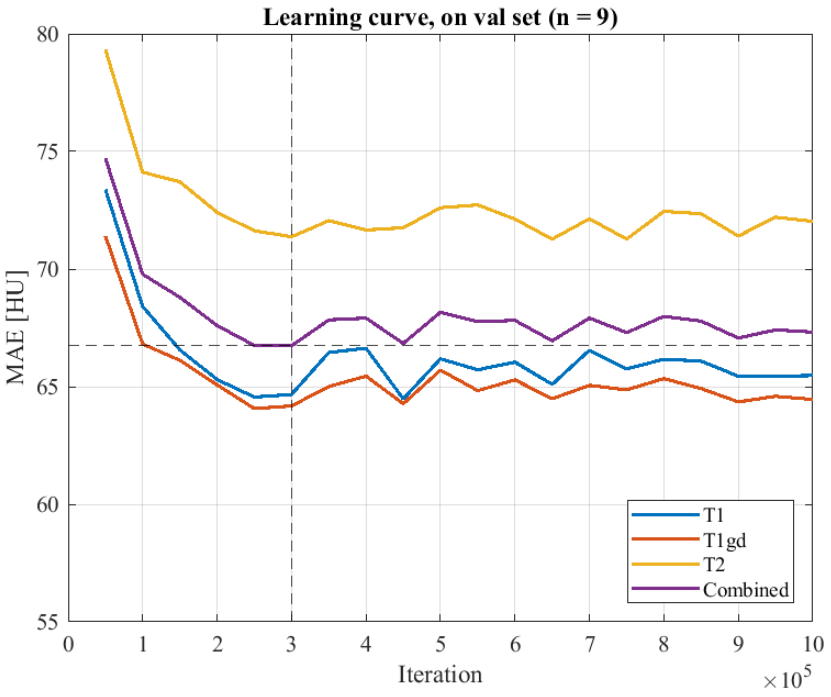

Fig. 22. Learning curves for the Baseline model on the validation set (n = 9) for each sequence separately: the MAE in the intersection of the body contour at each evaluated iteration is plotted against the iteration number. The combined MAE is the average over T1w(Gd) and T2w images. The first iteration for which the combined MAE did not improve for the following three iterations was selected. The evaluation was done every 50,000 iterations.

H. Image segmentation for creation of a training dataset of label maps: method

The methods for segmentation used to generate synthetic training data were briefly explained in section III-F.1. This section provides a more detailed explanation of the methods. Several methods were used to obtain an elaborate list of segmented structures (Table VII). Intracerebral structures were automatically segmented in T1w images using the open-source FastSurfer DL network [79]. The network requires input volumes of size [256, 256, 256]. Pre-processed T1w images were zero-padded if a dimension was less than 256 voxels, or zero-valued voxels were cropped from the volume in the case of larger dimensions to adhere to these sizes. After network inference, output label maps were reshaped to the original size of the T1w image. The segmentations of cerebrospinal fluid (CSF) and ventricles were grouped. OARs were added through automatic segmentation of T1w MRI using a previously in-house developed segmentation algorithm (unpublished) based on the DL model known as DeepMedic [81]. The model was previously developed for clinical use, employing the method as described in [80].

The GTV was added from an MRI-based clinical segmentation. Voxels in this GTV that had already been segmented as part of a cerebral structure with FastSurfer were assigned the corresponding FastSurfer label. Additionally, a clinical segmentation of the volume inside the skull was used to complement the segmentation of the CSF. All voxels falling inside the skull with no previous label (GTV, OAR, or from FastSurfer) were assigned the CSF label. A body contour was also obtained from the clinical segmentation. The intersection of this body contour and the registered MRI mask was used as a body mask for segmentation of some additional structures from CT_{train}.

Several structures were segmented using threshold operations on CT_{train}. Unless stated otherwise, thresholds were determined empirically based on one example patient from the training set and checked on two other patients. Background voxels and internal air were separated by applying a threshold of -0.5 to CT_{train}. The label for internal air was defined as those voxels falling inside the body mask but below the threshold. Segmentation of bone (i.e., bone and vertebrae) was achieved using a threshold of -0.120. Total bone was then subdivided into two classes (cortical bone and cancellous bone + bone marrow), using a threshold of 0.3216, with cortical bone defined as the voxels with an intensity above the threshold. The two bone labels were prioritised over the CSF label. Soft tissue was defined as all voxels inside the body mask that were not part of internal air and had not previously been assigned another label. Soft tissue was then divided into two classes (skin + muscle and other soft tissue) by thresholding the CT_{train} (-0.2078). This threshold was determined using Matlab's implementation of Otsu's method [97]: automatic multi-threshold computation with four thresholds was applied to one example patient. The most appropriate threshold was then chosen by manually checking the result.

TABLE VII

LOOK UP TABLE FOR THE AUTOMATICALLY SEGMENTED LABEL MAP AND METHODS USED FOR THE SEGMENTATION OF EACH STRUCTURE.

Label	**Structure**	**Segmentation method**
0	Background	Outside body mask (clinical segmentation, MRI mask)
1	Cortical white matter (L)	FastSurfer [79] label 2
2	Cortical white matter (R)	FastSurfer label 41
3	Cortical grey matter (L)	FastSurfer labels 1000-1999
4	Cortical grey matter (R)	FastSurfer labels $\geq$ 2000
5	Cerebellar white matter (L)	FastSurfer label 7
6	Cerebellar white matter (R)	FastSurfer label 46
7	Cerebellar cortex (L)	FastSurfer label 8
8	Cerebellar cortex (R)	FastSurfer label 47
9	Thalamus (L)	FastSurfer label 10
10	Thalamus (R)	FastSurfer label 49
11	Caudate nucleus (L)	FastSurfer label 11
12	Caudate nucleus (R)	FastSurfer label 50
13	Putamen (L)	FastSurfer label 12
14	Putamen (R)	FastSurfer label 51
15	Pallidum (L)	FastSurfer label 13
16	Pallidum (R)	FastSurfer label 52
17	Hippocampus (L)	FastSurfer label 17
18	Hippocampus (R)	FastSurfer label 53
19	Amygdala (L)	FastSurfer label 18
20	Amygdala (R)	FastSurfer label 54
21	Accumbens (L)	FastSurfer label 26
22	Accumbens (R)	FastSurfer label 58
23	Ventral diencephalon (L)	FastSurfer label 28
24	Ventral diencephalon (R)	FastSurfer label 60
25	Choroid plexus (L)	FastSurfer label 31
26	Choroid plexus (R)	FastSurfer label 63
27	White matter hypointensities (if exist)	FastSurfer label 77
28	Brain stem	FastSurfer label 16
29	CSF	FastSurfer labels 4, 5, 14, 15, 24, 43 and 44 and voxels inside skull (clinical segmentation) with no other label
30	GTV	Clinical segmentation, not labelled by FastSurfer
31	Eye (L)	DeepMedic-based OAR segmentation network
32	Eye (R)	DeepMedic-based OAR segmentation network
33	Lense (L)	DeepMedic-based OAR segmentation network
34	Lense (R)	DeepMedic-based OAR segmentation network
35	Cochlea (L)	DeepMedic-based OAR segmentation network
36	Cochlea (R)	DeepMedic-based OAR segmentation network
37	Lacrimal gland (L)	DeepMedic-based OAR segmentation network
38	Lacrimal gland (R)	DeepMedic-based OAR segmentation network
39	Optic nerve (L)	DeepMedic-based OAR segmentation network
40	Optic nerve (R)	DeepMedic-based OAR segmentation network
41	Pituitary gland	DeepMedic-based OAR segmentation network
42	Optic chiasm	DeepMedic-based OAR segmentation network
43	Bone: Cortical bone	Thresholding CT_{train}
44	Bone: cancellous bone + bone marrow	Thresholding CT_{train}
45	soft tissue: muscle + skin	Thresholding CT_{train}
46	soft tissue: other soft tissue	Thresholding CT_{train}
47	Internal air	Thresholding CT_{train}, falling inside body mask (clinical segmentation, MRI mask)

I. Domain randomisation - Random contrast: dataset balancing

Dataset balancing experiments were conducted for the two domain randomisation methods compared in section III-H.3. This section describes the dataset balancing experiment for the domain randomisation method using random contrast images.

A. Methods

For the domain randomisation method regarding RC images, dataset balancing was done by training two models on the whole training set (n = 60). The RC-only model was trained on RC images only, i.e., the training dataset of this model consisted of label maps only (section III-F.1; n = 60), which were converted to RC images on the fly. The training dataset of the RC+T1(Gd)+T2 model consisted of a mix of label maps (n = 60) and acquired T1w (n = 30), T1wGd (n = 30) and T2w (n = 60) images. Both models were trained using the model configuration and hyperparameters described in section III-E. For both models, early stopping was applied after 450,000 iterations, basing the decision on the MAE in sCT generated from T1w, T1wGd and T2w images as illustrated in Appendix G.

Image similarity metrics were computed for both models on the FLAIR images of the validation set for the chosen iteration. Performance was compared for all sequences: the model with the best performance of the two was chosen for comparison with the best model resulting from dataset balancing for the linear combination-based domain randomisation strategy (Appendix J). The MAE was leading in the model choice.

B. Results

The RC+T1(Gd)+T2 model outperformed the RC-only model on all sequences (Table VIII), with statistically significant differences in the MAE obtained for T1w, T1wGd and T2w images (p-values in Table IX). The most considerable difference in MAE was obtained for T2w images, finding values of 76.3 $\pm$ 10.9 HU and 128 $\pm$ 14.6 HU for the RC+T1(Gd)+T2 and RC-only model. For FLAIR images, values of 105 $\pm$ 20.5 HU and 109 $\pm$ 19.1 were found ($p > 0.05$). SSIM and PSNR are in line with the results for MAE, except the improvement in SSIM on FLAIR images for the RC+T1(Gd)+T2 model was statistically significant. Based on these findings, the RC+T1(Gd)+T2 model was chosen for comparison in section IV-A.3.

TABLE VIII

Image similarity metrics per MRI sequence for sCT generated by the RC-only and RC+T1(Gd)+T2 model, compared to ground truth CT.

Metric	**Model**	*Sequence*			
		T1w	**T1wGd**	**T2w**	**FLAIR**
MAE [HU]	RC-only	103 $\pm$ 15.7 [82.9 - 129]	100 $\pm$ 17.0 [75.6 - 134]	128 $\pm$ 14.6 [106 - 150]	109 $\pm$ 19.1 [74.9 - 136]
	RC+T1(Gd)+T2	71.5 $\pm$ 12.1 [59.7 - 100]	69.6 $\pm$ 12.2 [56.6 - 98.6]	76.3 $\pm$ 10.9 [60.2 - 95.6]	105 $\pm$ 20.5 [74.1 - 142]
SSIM	RC-only	0.796 $\pm$ 0.0403 [0.701 - 0.837]	0.803 $\pm$ 0.0436 [0.700 - 0.844]	0.747 $\pm$ 0.0321 [0.697 - 0.793]	0.782 $\pm$ 0.0527 [0.690 - 0.851]
	RC+T1(Gd)+T2	0.869 $\pm$ 0.0265 [0.801 - 0.889]	0.873 $\pm$ 0.0265 [0.804 - 0.895]	0.855 $\pm$ 0.024 [0.804 - 0.887]	0.803 $\pm$ 0.0467 [0.720 - 0.865]
PSNR [dB]	RC-only	25.4 $\pm$ 1.21 [23.9 - 27.2]	25.7 $\pm$ 1.36 [23.6 - 28.2]	23.7 $\pm$ 0.861 [22.4 - 25.2]	25.1 $\pm$ 1.51 [23.2 - 28.3]
	RC+T1(Gd)+T2	28.1 $\pm$ 1.25 [25.7 - 29.9]	28.3 $\pm$ 1.38 [25.6 - 30.4]	27.5 $\pm$ 1.15 [25.8 - 29.4]	25.6 $\pm$ 1.50 [23.7 - 28.5]

Metrics were calculated on the validation set (n = 9) within the intersection of the body contour of the sCT and CT. Mean values and standard deviations ($\mu \pm 1\sigma$) and range ([min - max]) are reported.

TABLE IX

Statistical comparisons between image similarity metrics obtained for the RC-only and RC+T1(Gd)+T2 model. Results as p-values per sequence.

Metric	*Sequence*			
	T1w	**T1wGd**	**T2w**	**FLAIR**
MAE	0.04	0.004	0.004	0.3
SSIM	0.004	0.004	0.004	0.03
PSNR	0.004	0.004	0.004	0.1

Metrics were calculated on the validation set (n = 9) within the intersection of the body contour of the sCT and CT. Wilcoxon-signed rank tests were performed for the computation of p-values.

Values of $p < 0.05$ were regarded as statistically significant.

J. Domain randomisation - Linear combinations: dataset balancing

Dataset balancing experiments were conducted for the two domain randomisation methods compared in section III-H.3. This section describes the dataset balancing experiment for the domain randomisation method using linear combination images.

A. Methods

Dataset balancing was performed by training two models on the whole training set (n = 60). The LC-only model was trained on LC images only: the chance of applying a linear combination was 1. The LC+T1(Gd)+T2 model was trained to study the effect of increasing the probability of sampling a fully T_1- or fully T_2-weighted image.

The dataset from which LC images were generated consisted of acquired T1w (n = 60) images and T1wGd (n = 60) and T2w (n = 60) images that had been registered to their T1w counterpart. For the LC+T1(Gd)+T2 model, this LC-specific dataset and the dataset of 30 T1w images, 30 T1wGd images and 60 T2w images from previous experiments were used. A random choice was made whether or not to apply a linear combination, with a 50 % chance of doing so. The original dataset was sampled if an LC image should not be used.

Both models were trained with the model configuration and hyperparameters described in section III-E. Early stopping was applied as illustrated in Appendix G, stopping after 250,000 and 200,000 iterations for the LC-only and LC+T1(Gd)+T2 model. Image similarity metrics were computed for both models on the FLAIR images of the validation set for the chosen iteration. The models' performances were compared, using the best model to compare with the model resulting from dataset balancing for the domain randomisation method based on RC images (the RC+T1(Gd)+T2 model; see Appendix J).

B. Results

For each sequence, a lower MAE was obtained for the LC+T1(Gd)+T2 model than for the LC-only model (Table X), although the differences were only statistically significant for T1wGd and T2w images (p-values in Table XI). For T1wGd images, an MAE of 74.8 $\pm$ 13.2 HU was obtained for the LC-only model, compared to an MAE of 71.0 $\pm$ 12.2 HU for the LC+T1(Gd)+T2 model. The MAE on T2w images improved from 82.3 $\pm$ 12.1 HU for the LC-only model to 77.8 $\pm$ 11.4 HU for the LC+T1(Gd)+T2 model. The results for PSNR are in line with the results for MAE. A conflicting result was found for SSIM on T2w images: contrary to the results found for the other metrics, the SSIM for T2w images was statistically significantly better for the LC-only model. Differences in SSIM for the other sequences were not significant. Altogether, the results favour the LC+T1(Gd)+T2 model. Therefore, this model was compared to the RC+T1(Gd)+T2 model in section IV-A.3.

TABLE X

IMAGE SIMILARITY METRICS PER MRI SEQUENCE FOR SCT GENERATED BY THE LC-ONLY AND LC+T1(GD)+T2 MODEL, COMPARED TO GROUND TRUTH CT.

Metric	**Model**	*Sequence*			
		T1w	**T1wGd**	**T2w**	**FLAIR**
MAE [HU]	LC-only	75.0 $\pm$ 14.7 [62.1 - 112]	74.8 $\pm$ 13.2 [64.3 - 108]	82.3 $\pm$ 12.1 [68.2 - 106]	115 $\pm$ 27.8 [77.7 - 166]
	LC+T1(Gd)+T2	72.3 $\pm$ 12.4 [57.3 - 100]	71.0 $\pm$ 12.2 [58.2 - 99.8]	77.8 $\pm$ 11.4 [63.0 - 100]	110 $\pm$ 23.9 [72.9 - 155]
SSIM	LC-only	0.865 $\pm$ 0.0299 [0.787 - 0.882]	0.867 $\pm$ 0.0266 [0.798 - 0.886]	0.859 $\pm$ 0.0275 [0.799 - 0.896]	0.788 $\pm$ 0.0542 [0.695 - 0.860]
	LC+T1(Gd)+T2	0.870 $\pm$ 0.0255 [0.806 - 0.893]	0.871 $\pm$ 0.0257 [0.805 - 0.891]	0.854 $\pm$ 0.0247 [0.799 - 0.880]	0.793 $\pm$ 0.0474 [0.709 - 0.864]
PSNR [dB]	LC-only	27.9 $\pm$ 1.38 [24.9 - 29.7]	27.8 $\pm$ 1.28 [25.1 - 29.5]	26.9 $\pm$ 1.09 [25.2 - 28.6]	25.0 $\pm$ 1.85 [22.6 - 28.2]
	LC+T1(Gd)+T2	28.1 $\pm$ 1.33 [25.8 - 30.3]	28.2 $\pm$ 1.35 [25.6 - 30.2]	27.4 $\pm$ 1.14 [25.7 - 29.2]	25.3 $\pm$ 1.72 [23.0 - 28.7]

Metrics were calculated on the validation set (n = 9) within the intersection of the body contour of the sCT and CT. Mean values and standard deviations ($\mu \pm 1\sigma$) and range ([min - max]) are reported.

TABLE XI

STATISTICAL COMPARISONS BETWEEN IMAGE SIMILARITY METRICS OBTAINED FOR THE LC-ONLY AND LC+T1(GD)+T2 MODEL. RESULTS AS P-VALUES PER SEQUENCE.

Metric	*Sequence*			
	T1w	**T1wGd**	**T2w**	**FLAIR**
MAE	0.1	0.004	0.008	0.055
SSIM	0.2	0.1	0.02	0.1
PSNR	0.1	0.004	0.004	0.055

Metrics were calculated on the validation set (n = 9) within the intersection of the body contour of the sCT and CT. Wilcoxon-signed rank tests were performed for the computation of p-values.

Values of $p < 0.05$ were regarded as statistically significant.

K. Experiments: image similarity metrics calculated per sequence for models presented in experiments

As a complement to the mean values and standard deviations for MAE presented in section IV-A, Table XII presents results for MAE, SSIM and PSNR per model trained in the experiments. The range is presented in addition to the mean value and standard deviation.

TABLE XII
IMAGE SIMILARITY METRICS PER MRI SEQUENCE FOR SCT GENERATED BY MODELS PRESENTED IN THE EXPERIMENTS, COMPARED TO GROUND TRUTH CT.

Metric	**Model**	*Sequence*			
		T1w	**T1wGd**	**T2w**	**FLAIR**
MAE [HU]	Baseline	68.4 ± 14.2 [55.1 - 103]	67.8 ± 13.7 [55.7 - 101]	74.7 ± 13.3 [55.2 - 100]	114 ± 28.4 [75.6 - 172]
	T1-only	67.2 ± 14.2 [54.3 - 101]	66.2 ± 14.7 [53.5 - 102]	136 ± 21.1 [106 - 164]	125 ± 31.6 [80.4 - 183]
	RC+T1(Gd)	69.1 ± 12.6 [57.3 - 99.5]	68.2 ± 13.2 [56.7 - 99.4]	111 ± 14.6 [92.4 - 133]	100 ± 18.5 [71.0 - 132]
	RC+T1(Gd)+T2	71.5 ± 12.1 [59.7 - 100]	69.6 ± 12.2 [56.6 - 98.6]	76.3 ± 10.9 [60.2 - 95.6]	105 ± 20.5 [74.1 - 142]
	LC+T1(Gd)+T2	72.3 ± 12.4 [57.3 - 100]	71.0 ± 12.2 [58.2 - 99.8]	77.8 ± 11.4 [63.0 - 100]	110 ± 23.9 [72.9 - 155]
SSIM	Baseline	0.877 ± 0.0283 [0.806 - 0.899]	0.877 ± 0.0282 [0.805 - 0.898]	0.860 ± 0.0284 [0.798 - 0.896]	0.792 ± 0.0516 [0.696 - 0.865]
	T1-only	0.880 ± 0.0278 [0.811 - 0.902]	0.880 ± 0.0298 [0.804 - 0.902]	0.747 ± 0.0399 [0.700 - 0.821]	0.771 ± 0.0589 [0.681 - 0.856]
	RC+T1(Gd)	0.876 ± 0.0268 [0.807 - 0.893]	0.877 ± 0.0269 [0.809 - 0.987]	0.785 ± 0.0317 [0.731 - 0.835]	0.811 ± 0.0447 [0.729 - 0.870]
	RC+T1(Gd)+T2	0.869 ± 0.0265 [0.801 - 0.889]	0.873 ± 0.0265 [0.804 - 0.895]	0.855 ± 0.024 [0.804 - 0.887]	0.803 ± 0.0467 [0.720 - 0.865]
	LC+T1(Gd)+T2	0.870 ± 0.0255 [0.806 - 0.893]	0.871 ± 0.0257 [0.805 - 0.891]	0.854 ± 0.0247 [0.799 - 0.880]	0.793 ± 0.0474 [0.709 - 0.864]
PSNR [dB]	Baseline	28.5 ± 1.56 [25.3 - 30.5]	28.5 ± 1.63 [25.2 - 30.5]	27.7 ± 1.41 [25.6 - 30.2]	25.2 ± 1.89 [22.3 - 28.7]
	T1-only	28.5 ± 1.54 [25.4 - 30.6]	28.6 ± 1.70 [25.0 - 30.7]	23.3 ± 1.09 [21.7 - 24.9]	24.6 ± 1.92 [22.0 - 28.1]
	RC+T1(Gd)	28.4 ± 1.35 [25.6 - 30.1]	28.5 ± 1.52 [25.4 - 30.4]	24.8 ± 0.967 [23.3 - 26.3]	25.9 ± 1.50 [24.0 - 28.8]
	RC+T1(Gd)+T2	28.1 ± 1.25 [25.7 - 29.9]	28.3 ± 1.38 [25.6 - 30.4]	27.5 ± 1.15 [25.8 - 29.4]	25.6 ± 1.50 [23.7 - 28.5]
	LC+T1(Gd)+T2	28.1 ± 1.33 [25.8 - 30.3]	28.2 ± 1.35 [25.6 - 30.2]	27.4 ± 1.14 [25.7 - 29.2]	25.3 ± 1.72 [23.0 - 28.7]

Metrics were calculated on the validation set (n = 9) within the intersection of the body contour of the sCT and CT. Mean values and standard deviations ($\mu \pm 1\sigma$) and range ([min - max]) are reported.

L. Results: p-values obtained with statistical comparisons in the experiments and final comparison

This section provides the p-values obtained with Wilcoxon-signed rank tests conducted for comparisons presented in the body of this work, both for image similarity metrics (experiments and final comparison) and for dosimetric accuracy (final comparison only). In the experiments, statistical comparisons of image similarity metrics for pairs of models were made per sequence (Table XIII), with metrics calculated on the validation set (n = 9). Similarly, image similarity metrics obtained on the test set (n = 25) per sequence were compared for pairs of models included in the final comparison (Table XIV). For the models in the final comparisons, p-values obtained with statistical comparisons are shown in the violin plots in the body of this work.

TABLE XIII

STATISTICAL COMPARISONS BETWEEN IMAGE SIMILARITY METRICS FOR PAIRS OF MODELS THAT WERE COMPARED IN THE EXPERIMENTS. RESULTS AS P-VALUES PER SEQUENCE.

Metric	**Compared models**	*Sequence*			
		T1w	**T1wGd**	**T2w**	**FLAIR**
MAE	Baseline vs T1-only	0.2	0.04	0.004	0.004
	RC+T1(Gd) vs T1-only	0.1	0.055	0.004	0.004
	RC+T1(Gd) vs Baseline	0.3	0.7	0.004	0.004
	RC+T1(Gd)+T2 vs LC+T1(Gd)+T2	0.3	0.2	0.2	0.04
SSIM	Baseline vs T1-only	0.01	0.07	0.004	0.004
	RC+T1(Gd) vs T1-only	0.004	0.07	0.004	0.004
	RC+T1(Gd) vs Baseline	0.3	0.5	0.004	0.004
	RC+T1(Gd)+T2 vs LC+T1(Gd)+T2	0.9	0.4	0.4	0.07
PSNR	Baseline vs T1-only	0.5	0.2	0.004	0.004
	RC+T1(Gd) vs T1-only	0.4	0.3	0.004	0.004
	RC+T1(Gd) vs Baseline	0.4	0.6	0.004	0.004
	RC+T1(Gd)+T2 vs LC+T1(Gd)+T2	1	0.6	0.4	0.07

Metrics were calculated on the validation set (n = 9) within the intersection of the body contour of the sCT and CT. Wilcoxon-signed rank tests were done for the computation of p-values.

Values of $p < 0.05$ were regarded as statistically significant.

TABLE XIV

STATISTICAL COMPARISONS OF IMAGE SIMILARITY METRICS BETWEEN MODELS FOR BASELINE, BASELINE+FLAIR AND DOMAIN RANDOMISATION MODEL. RESULTS AS P-VALUES PER SEQUENCE.

Metric	**Compared models**	*Sequence*			
		T1w	**T1wGd**	**T2w**	**FLAIR**
MAE	Baseline vs Baseline+FLAIR	0.007	0.1	$3 * 10^{-4}$	$1 * 10^{-5}$
	Baseline vs Domain Randomisation	$2 * 10^{-4}$	$3 * 10^{-5}$	$2 * 10^{-4}$	$3 * 10^{-5}$
	Baseline+FLAIR vs Domain Randomisation	$4 * 10^{-4}$	$5 * 10^{-4}$	0.4	$1 * 10^{-5}$
SSIM	Baseline vs Baseline+FLAIR	0.054	0.3	$5 * 10^{-}4$	$1 * 10^{-5}$
	Baseline vs Domain Randomisation	0.002	$2 * 10^{-4}$	0.002	$3 * 10^{-4}$
	Baseline+FLAIR vs Domain Randomisation	0.003	$9 * 10^{-4}$	0.99	$1 * 10^{-5}$
PSNR	Baseline vs Baseline+FLAIR	0.007	0.07	$4 * 10^{-4}$	$1 * 10^{-5}$
	Baseline vs Domain Randomisation	$7 * 10^{-4}$	$2 * 10^{-4}$	0.001	0.002
	Baseline+FLAIR vs Domain Randomisation	0.002	0.007	0.6	$1 * 10^{-5}$

Metrics were calculated on the test set (n = 25) within the intersection of the body contour of the sCT and CT. P-values were calculated with Wilcoxon-signed rank tests.

Values of $p < 0.05$ were regarded as statistically significant.

For each model included in the final comparison, dosimetric accuracy was statistically compared between sequences (Table XV). Additionally, comparisons were made in dosimetric accuracy between pairs of models, with p-values reported per sequence (Table XVI).

TABLE XV

STATISTICAL COMPARISONS OF METRICS FOR DOSIMETRIC ACCURACY BETWEEN SEQUENCES FOR BASELINE, BASELINE+FLAIR AND DOMAIN RANDOMISATION MODEL. RESULTS AS P-VALUES PER MODEL.

Metric	**Compared sequences**	*Model*		
		Baseline	**Baseline+FLAIR**	**Domain Randomisation**
$\gamma_{3\%,3mm}$	T1 vs T1gd	0.6	0.9	0.4
	T1 vs T2	0.8	0.97	1
	T1 vs FLAIR	0.02	0.1	0.4
	T1gd vs T2	0.6	1	0.7
	T1gd vs FLAIR	0.01	0.3	0.1
	T2 vs FLAIR	0.1	0.08	0.8
$\gamma_{2\%,2mm}$	T1 vs T1gd	0.2	0.4	0.98
	T1 vs T2	0.04	0.06	0.2
	T1 vs FLAIR	0.001	0.2	0.005
	T1gd vs T2	0.3	0.3	0.1
	T1gd vs FLAIR	0.005	0.2	0.01
	T2 vs FLAIR	0.1	0.5	0.2
$\gamma_{1\%,1mm}$	T1 vs T1gd	0.5	0.8	0.3
	T1 vs T2	0.7	0.04	0.3
	T1 vs FLAIR	$5 * 10^{-5}$	0.03	$4 * 10^{-4}$
	T1gd vs T2	0.4	0.1	0.09
	T1gd vs FLAIR	$1 * 10^{-5}$	0.001	$1 * 10^{-4}$;
	T2 vs FLAIR	$1 * 10^{-4}$	0.9	0.01
DD	T1 vs T1gd	0.2	0.4	0.2
	T1 vs T2	0.4	0.9	0.2
	T1 vs FLAIR	$2 * 10^{-4}$	0.2	$2 * 10^{-5}$
	T1gd vs T2	0.8	0.5	0.2
	T1gd vs FLAIR	$2 * 10^{-4}$	0.8	$3 * 10^{-5}$
	T2 vs FLAIR	$4 * 10^{-4}$	0.4	$1 * 10^{-5}$

Metrics were calculated on the test set (n = 25) within the intersection of the body contour of the sCT and CT. P-values were calculated with Wilcoxon-signed rank tests.

Values of $p < 0.05$ were regarded as statistically significant.

TABLE XVI

STATISTICAL COMPARISONS OF METRICS FOR DOSIMETRIC ACCURACY BETWEEN MODELS FOR BASELINE, BASELINE+FLAIR AND DOMAIN RANDOMISATION MODEL. RESULTS AS P-VALUES PER SEQUENCE.

Metric	**Compared models**	*Sequence*			
		T1w	**T1wGd**	**T2w**	**FLAIR**
$\gamma_{3\%,3mm}$	Baseline vs Baseline+FLAIR	0.7	0.97	0.9	0.003
	Baseline vs Domain Randomisation	0.1	0.7	0.7	0.07
	Baseline+FLAIR vs Domain Randomisation	0.04	0.4	0.8	0.006
$\gamma_{2\%,2mm}$	Baseline vs Baseline+FLAIR	0.8	0.2	0.5	0.01
	Baseline vs Domain Randomisation	0.4	0.7	0.6	0.3
	Baseline+FLAIR vs Domain Randomisation	0.7	0.4	0.9	0.053
$\gamma_{1\%,1mm}$	Baseline vs Baseline+FLAIR	0.003	0.1	0.2	$2 * 10^{-4}$
	Baseline vs Domain Randomisation	0.7	0.5	0.5	0.005
	Baseline+FLAIR vs Domain Randomisation	0.04	0.07	0.07	$9 * 10^{-5}$
DD [%]	Baseline vs Baseline+FLAIR	0.01	$7 * 10^{-4}$	0.002	$4 * 10^{-5}$
	Baseline vs Domain Randomisation	0.3	0.1	0.06	0.06
	Baseline+FLAIR vs Domain Randomisation	0.003	$2 * 10^{-4}$	$4 * 10^{-5}$	$1 * 10^{-4}$

Metrics were calculated on the test set (n = 25) within the intersection of the body contour of the sCT and CT. P-values were calculated with Wilcoxon-signed rank tests.

Values of $p < 0.05$ were regarded as statistically significant.

M. FINAL COMPARISON: DOSIMETRIC ACCURACY

As a complement to the results for dose accuracy (3D γ-pass rate with 1%,1mm criterion and DD in the high dose region) presented in section IV-C, Table XVII presents results for the γ-pass rates with 3%,3mm and 2%,2mm criteria.

TABLE XVII
DOSE EVALUATION ($\gamma_{3\%,3mm}$ AND $\gamma_{2\%,2mm}$) FOR SCT GENERATED BY THE BASELINE, BASELINE+FLAIR AND DOMAIN RANDOMISATION MODELS PER MRI SEQUENCE.

Metric	Model	*Sequence*			
		T1w	**T1wGd**	**T2w**	**FLAIR**
$\gamma_{3\%,3mm}$ [%] [a]	Baseline	99.99 ± 0.01 [99.96 - 100]	100 ± 0.01 [99.9 - 100]	99.99 ± 0.01 [99.97 - 100]	99.99 ± 0.02 [99.9 - 100]
	Baseline+FLAIR	100 ± 0.01 [99.95 - 100]	100 ± 0.01 [99.97 - 100]	99.99 ± 0.01 [99.96 - 100]	100 ± 0.01 [99.97 - 100]
	Domain Randomisation	99.99 ± 0.01 [99.95 - 100]	99.99 ± 0.01 [99.95 - 100]	99.99 ± 0.01 [99.96 - 100]	99.99 ± 0.01 [99.95 - 100]
$\gamma_{2\%,2mm}$ [%] [a]	Baseline	99.95 ± 0.1 [99.7 - 100]	99.95 ± 0.1 [99.6 - 100]	99.9 ± 0.1 [99.7 - 100]	99.9 ± 0.2 [99.4 - 100]
	Baseline+FLAIR	99.96 ± 0.1 [99.7 - 100]	99.95 ± 0.1 [99.7 - 100]	99.9 ± 0.1 [99.7 - 100]	99.95 ± 0.1 [99.7 - 100]
	Domain Randomisation	99.95 ± 0.08 [99.7 - 100]	99.95 ± 0.1 [99.6 - 100]	99.9 ± 0.08 [99.7 - 100]	99.9 ± 0.1 [99.5 - 100]

Dosimetric accuracy was assessed through plan re-calculation on water-filled sCT compared to the water-filled acquired CT. Mean values and standard deviations ($\mu \pm 1\sigma$) and range ([min - max]) are reported. [a]Calculated in the D $>$ 10 % prescribed region. [b]Calculated in the D $>$ 90 % prescribed region.

As part of the dosimetric evaluation of the generated sCT, DVH differences in D_{median} and D_{max} between sCT- and CT-based dose plans were evaluated for the brainstem, optic chiasm, lenses, cochleae and pituitary gland. Boxplots representing these differences are shown per sequence in Fig. 23 (Baseline model), Fig. 24 (Baseline+FLAIR model) and Fig. 25 (Domain Randomisation model). Section IV-C discusses the main differences.

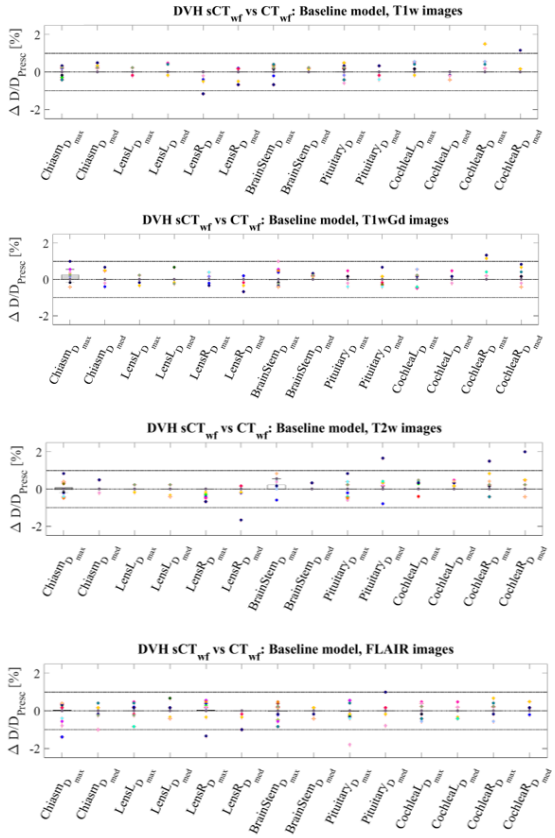

Fig. 23. Boxplots for DVH differences between sCT_{wf} generated by the Baseline model and CT_{wf}-based dose plans in D_{max} and D_{median} for OARs: brainstem, optic chiasm, lenses, cochleae and pituitary gland. Results are shown per sequence (top to bottom: T1w, T1wGd, T2w and FLAIR images). Dots represent outliers, with each colour representing a different patient.

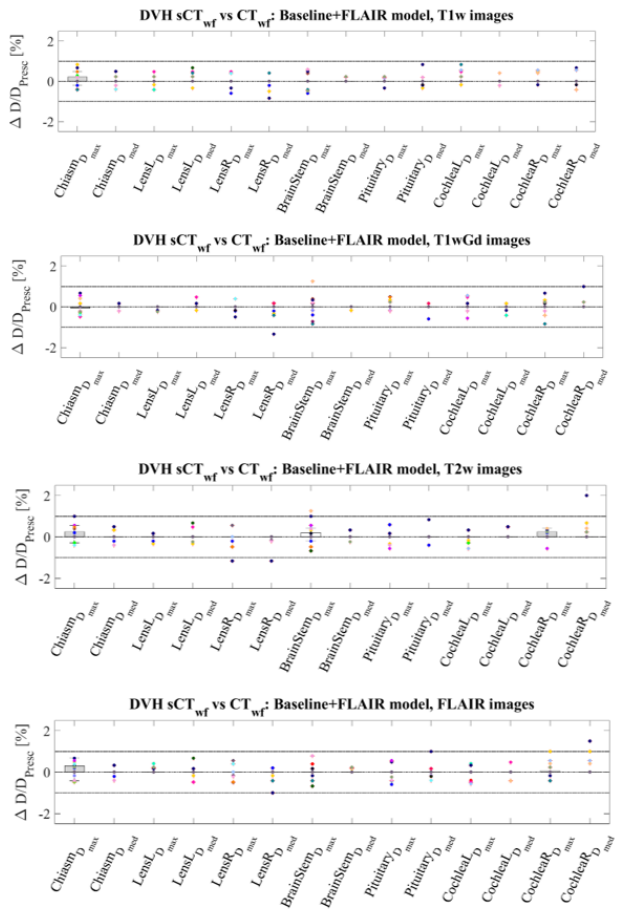

Fig. 24. Boxplots for DVH differences between sCT$_{wf}$ generated by the Baseline+FLAIR model and CT$_{wf}$-based dose plans in D_{max} and D_{median} for OARs: brainstem, optic chiasm, lenses, cochleae and pituitary gland. Results are shown per sequence (top to bottom: T1w, T1wGd, T2w and FLAIR images). Dots represent outliers, with each colour representing a different patient.

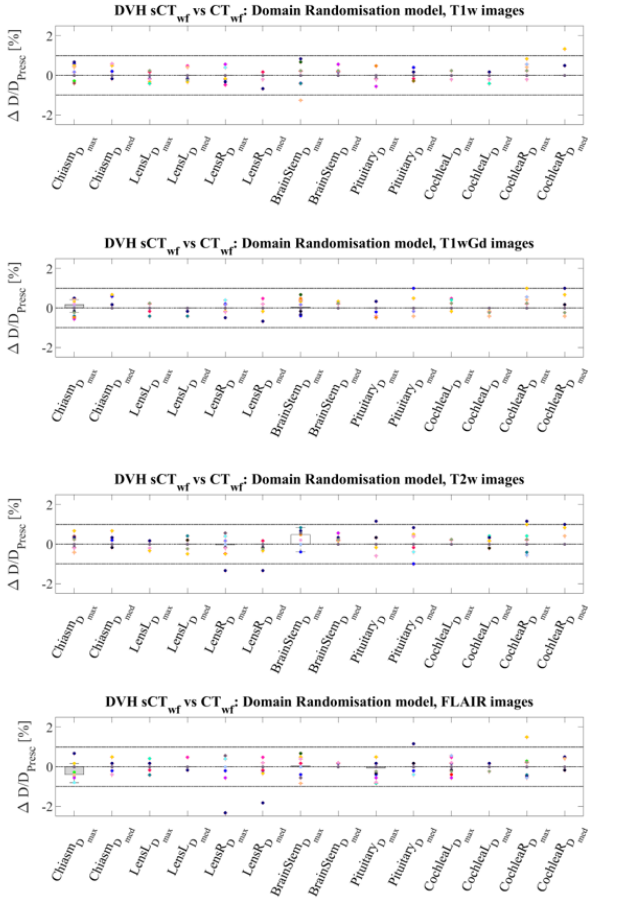

Fig. 25. Boxplots for DVH differences between sCT$_{wf}$ generated by the Domain Randomisation model and CT$_{wf}$-based dose plans in D_{max} and D_{median} for OARs: brainstem, optic chiasm, lenses, cochleae and pituitary gland. Results are shown per sequence (top to bottom: T1w, T1wGd, T2w and FLAIR images). Dots represent outliers, with each colour representing a different patient.

N. Search strategy for previous literature about contrast-agnostic deep learning-based MRI-to-CT synthesis

This section provides the search strategy adopted to verify that no previous work explored whether a single cGAN network can be trained for sCT generation from multiple MRI sequences without (re)training the network on new, unseen sequences.

A comprehensive search for studies published in the scientific database Scopus was performed on April 29th, 2022. The search strategy included the following search terms: (sCT OR ”synth* CT” OR ”CT synth*” OR pseudoCT OR ”pseudo CT” OR ”pseudo-CT” OR pCT OR ”MRI-to-CT”) AND (generalis* OR generaliz* OR ”contrast-agnostic” OR ”contrast-agnostic” OR sequence OR contrast) AND (GAN OR ”generative adversarial net*” OR ”deep learning” OR CNN OR Unet OR U-Net OR ”neural net*”). The search was performed on title, abstract and keywords with no limitation on the publication date or language.

O. Comparison to literature: sCT synthesis from brain MRI

A. Search strategy

On April 29th, 2022, a systematic literature search was done in the scientific database Scopus for studies about the accuracy of DL-based sCT generation from brain MRI for MR-only RT planning. The search strategy included the following keywords: (sCT OR "synth* CT" OR "CT synth*" OR pseudoCT OR "pseudo CT" OR "pseudo-CT" OR pCT OR "MRI-to-CT") AND (brain OR cerebr* OR head OR skull) and (GAN OR "generative adversarial net*" OR "deep learning" OR CNN OR Unet OR U-Net OR "neural net*"). The search was performed on title, abstract and keywords with no limitation on the publication date or language. The titles and abstracts were reviewed to select studies for full-text review. Any doubts about inclusion were resolved by screening the full text.

The MAE obtained from the comparison of sCT to acquired CT had to be reported, possibly with additional metrics SSIM, PSNR and γ-pass rates with 3%,3mm, 2%,2mm or 1%1,mm criterion. Only studies aiming to generate sCT from brain MRI scans for MR-only RT planning were considered for inclusion. Articles were excluded if: a. the full text was unavailable in Dutch or English; b. the article was a conference paper or review; c. the aim was CT-to-MRI translation instead of MRI-to-CT translation; d. patients were solely head-and-neck cancer patients instead of brain cancer patients; e. either the input sequence was not a T1w, T1w+Gd, T2w or T2w FLAIR image, or the input sequence was undefined; or f. the article was a duplicate evaluation of a DL model already evaluated in an earlier study.

B. Data extraction

A spreadsheet was designed for data extraction, extracting the following information from the included articles: a. basic information, including the first author to allow identification, year of publication, journal; and b. data needed for comparison with results obtained in the current work: input MRI sequence, model configuration, type of model, image similarity metrics (MAE, SSIM, PSNR), and γ-pass rates for 3%,3mm, 2%,2mm and 1%,1mm criteria. Dose differences were not considered because of the large variability in reported metrics.

C. Results

Only three studies were identified that presented results for T2w FLAIR images separately. Most included studies (n = 17; Table XVIII) used T1w images as input sequence, followed by T1wGd images (n = 7) and T2w images (n = 5). The MAE obtained for models taking T1w images as input sequences ranged between 45.4 HU [83] to 131 HU [84], in addition to one exceptionally low MAE of 9.02 HU [98] (Table XVIII). For T1wGd images, mean MAEs ranged from 44.6 HU [83] to 89.3 HU [85]. Values between 45.7 HU [83] and 68.3 HU [57] were identified for T2w images. For T2w FLAIR images, two identified studies reported MAE values of 51.2 HU [83] and 59.3 HU [87], and one study reported an MAE of 115 $\pm$ 22. Additionally, in [99], a mean MAE of 61.9 $\pm$ 22.6 was obtained for a model trained on a mix of T2w images with and without FLAIR. However, no statistical comparisons were provided between groups of patients with different imaging protocols [99].

TABLE XVIII
LITERATURE REVIEW: BRAIN sCT WITH IMAGE SIMILARITY METRICS AND GAMMA ANALYSIS.

Training sequence	Reference	Pts[a]	Conf.[b]	Model	Image similarity			Gamma analysis		
					MAE [HU]	SSIM	PSNR [dB]	$\gamma_{3\%,3mm}$ [%]	$\gamma_{2\%,2mm}$ [%]	$\gamma_{1\%,1mm}$ [%]
T1w	Han 2017 [100]	18	2D	U-net	84.8±17.3	-	-	-	-	-
	Dinkla 2018 [45]	26	2D+[c]	CNN	67±11	-	-	99.9±0.2	99.1±0.80	97.0±2.2
	Xiang 2018 [101]	16	2.5D[d]	U-net	85.4±9.24	-	27.3±1.1	-	-	-
	Gupta 2019 [102]	47	2D	U-net	81.0±14.6	-	-	-	-	-
	Koike 2019 [103]	15	2Dp	GAN	120±20.4	-	-	99.7±0.5	98.7±1.2	94.2±4.9
	Lei 2019 [104]	24	3Dp	GAN	55.7±9.4	-	25.8±1.81	-	-	-
	Neppl 2019 [95]	57	2D	U-net	116±26	-	-	-	98±2	-
	Shafai-Erfani 2019 [105]	25	3Dp	GAN	54.6±6.81	-	-	99.96±0.21	98.4±3.51	90.8±7.8
	Spadea 2019 [31]	12	2D+[e]	U-net	54±7	-	-	-	-	-
	Alvarez-Andres 2020 [86][f]	134	3Dp	CNN	84±25	-	-	99.8±0.18	99.6±0.33	97.9±1.16
	Massa 2020 [83]	81	2D	U-net	45.4±8.52	0.65±0.05	43.0±2.02	-	-	-
	Xu 2020 [98]	33	2D	GAN	9.02±0.82	0.75±0.77	-	-	-	-
	Irmak 2021 [84]	20	2D	GAN	131±14.3	-	-	-	99.0±0.4	95.2±1.9
	Sreeja 2021 [106]	19	2D	U-net	67.5±17.3	0.86±0.05	-	-	-	-
	Tang 2021 [107]	27	2D	GAN	60.8±14.0	-	49.23±1.92	99.96	98.0	-
	Zimmermann 2021 [57][f]	33	3D	U-net	68.1±5.4	0.97±0.00	-	-	-	-
	Gholamiankhah 2022 [108]	86	2D	CNN	114±27.5	0.95±0.04	28.7±1.59	-	-	-
	Wang 2022 [87][g]	145	2D	GAN	50.2±18	0.92±0.03	31.8±2.6	-	-	-
T1wGd	Emami 2018 [85] and Liu, 2021 [109][h]	15	2D	GAN	89.3±10.3	0.83±0.03	26.6±1.2	-	99.9±0.2	99.0±1.5
	Kazemifar 2019 [110]	63	2D	GAN	47.2±11.0	-	-	99.2±0.8	94.6±2.9	
	Liu, 2019 [111]	40	2D	CNN	75±23	-	-	99.2	-	
	Alvarez-Andres 2020 [86][f]	133	3Dp	CNN	87±28	-	-	99.9±0.18	99.6±0.30	97.9±1.07
	Massa 2020 [83]	81	2D	U-net	44.6±7.48	0.64±0.03	43.4±1.22	-	-	-
	Li 2021 [56]	18	2D	GAN	74.9±15.6	0.83±0.04	27.7±1.43	-	-	-
	Zimmermann 2021 [57][f]	24	3D	U-net	71.6±9.4	0.96±0.01	-	-	-	-
T1w + T1wGd	Alvarez-Andres 2020 [86][f]	242	3Dp	CNN	81±22	-	-	99.8±0.19	99.6±0.32	97.9±1.06
	Maspero 2020 [63][g]	40	2D+[e]	GAN	61.0±14.1	-	26.7±1.9	99.7 ±0.6	99.6±1.1	-
	Jabbarpour 2022 [99][i]	60	2D	GAN	62.7±30.7	0.88±0.05	27.0±3.38	99.0±1.10	95.0±3.68	90.1±6.05
T2wGd	Li 2020 [112]	28	2D	U-net	65.4±4.08	0.97±0.004	28.84±0.57	-	-	-
	Massa 2020 [83]	81	2D	U-net	45.7±8.78	0.63±0.03	43.4±1.18	-	-	-
	Ranjan 2021 [113]	18	2D	GAN	0.03±0.02[j]	0.82±0.06	21.4±3.96	-	-	-
	Zimmermann 2021 [57][f]	32	3D	U-net	68.3±7.3	0.98±0.00	-	-	-	-
	Wang 2022 [87][g]	145	2D	GAN	53.7±21	0.91±0.02	32.5±2.2	-	-	-
T1w + T1wGd + T2w	Zimmermann 2021 [57][f]	33	3D	U-net	T1w: 69.1±5.7 T1wGd: 70.0±8.4 T2w: 67.3±7.1	T1w: 0.97±0.00 T1wGd: 0.97±0.01 T2w: 0.98±0.00	-	-	-	-
FLAIR	Alvarez-Andres 2020 [86][f]	134	3Dp	CNN	115±22	-	-	-	-	-
	Massa 2020 [83]	81	2D	U-net	51.2±4.5	0.61±0.04	44.9±1.15	-	-	-
	Wang 2022 [87][g]	145	2D	GAN	59.3±22	0.91±0.03	31.3±2.0	-	-	-
T2w + FLAIR	Jabbarpour 2022 [99][i]	65	2D	GAN	61.9±22.6	0.84±0.05	27.1±2.25	-	-	-
Multichannel: T1w + T2w + FLAIR	Koike 2019 [103]	15	2Dp	GAN	108±24.0	-	-	99.8±0.3	99.2±1.0	95.3±4.7

For references where multiple models were compared, results are only reported for the best performing model unless models were trained for different MRI sequences. [a]Number of patients in the training set. [b]Configuration, p: patch-based training. [c]Three orthogonal slices as input. [d]Three consecutive slices as input. [e]Combined output from three networks that take one direction from the three orthogonal planes as input. [f]Results reported for several single-sequence models and a combined model. [g]Paediatric population. [h]Dosimetry reported in [109] for the model presented in [85]. [i]Heterogeneous imaging protocol, incl. images +/- Gd and +/- FLAIR. [j]MAE not in HU.

P. Example sCT images illustrating possible effects of data imperfections on network performance

This section contains two images illustrating the effects of an imperfection in training and test data on model performance, as mentioned in V.

Partial volume effects were observed for T2w images, which may have influenced the performance of all models trained in this work for this specific sequence. For instance, in the illustrative T2w image from the test set shown in Fig. 26 (left), the border between the skull and surrounding tissues is blurred, causing the skull to be mapped too thick in the sCT generated by the Baseline model in this area (right).

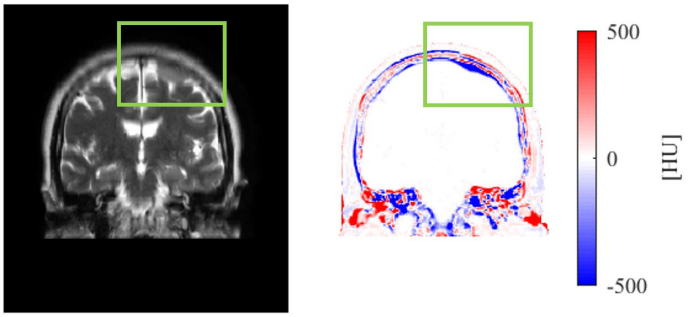

Fig. 26. T2w image (left) of PT16 for whom partial volume effects can be observed near the border of the skull (rectangle). The image on the right shows the difference between the corresponding sCT generated by the Baseline model and the acquired CT, with a discrepancy between the two in the same area (rectangle).

Figure 27 illustrates how a possible error registering the acquired MRI to the corresponding CT might affect the MAE. The transversal slice shown suggests the input MRI was rotated to the corresponding ground truth CT. If such misregistrations occur in the training data, network performance might be negatively influenced [88]. Likewise, suppose the test data are not correctly registered. In that case, the networks might be penalised for discrepancies between sCT and acquired CT that are, in reality, caused by errors in registration and not by improper mapping from MRI to CT.

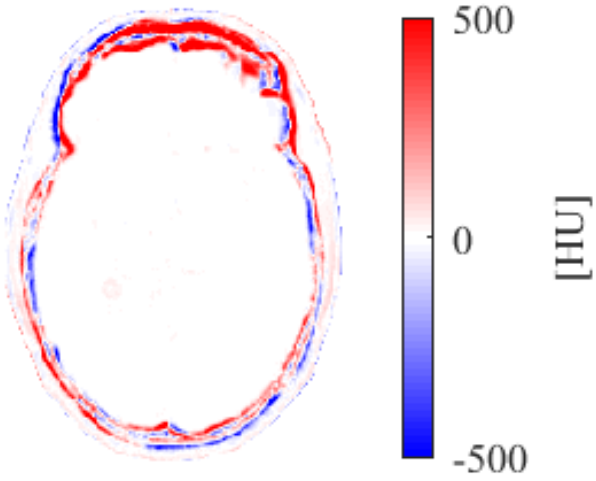

Fig. 27. Difference between acquired CT and the sCT produced by the Baseline model from a T1w image for an example patient showing the effect of misregistration: the transversal slice suggests the input MRI was rotated to the corresponding ground truth CT. Similar results were obtained for this patient for the other models.