This is the published version of a paper published in *Physics in Medicine and Biology*.

Citation for the original published paper (version of record):

Fetty, L., Löfstedt, T., Heilemann, G., Furtado, H., Nesvacil, N. et al. (2020) Investigating conditional GAN performance with different generator architectures, an ensemble model, and different MR scanners for MR-sCT conversion *Physics in Medicine and Biology*, 65(10): 105004 https://doi.org/10.1088/1361-6560/ab857b

Access to the published version may require subscription.

N.B. When citing this work, cite the original published paper.

Permanent link to this version:

http://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-172540
Investigating conditional GAN performance with different generator architectures, an ensemble model, and different MR scanners for MR-sCT conversion

To cite this article: Lukas Fetty et al 2020 Phys. Med. Biol. 65 105004

View the article online for updates and enhancements.
Investigating conditional GAN performance with different generator architectures, an ensemble model, and different MR scanners for MR-sCT conversion

Lukas Fetty, Tommy Löfstedt, Gerd Heilemann, Hugo Furtado, Nicole Nesvacil, Tufve Nyholm, Dietmar Georg, Peter Kuess

1 Division of Medical Radiation Physics, Department of Radiation Oncology, Medical University of Vienna, Vienna, Austria
2 Medical Imaging Cluster, Medical University of Vienna, Vienna, Austria
3 Department of Radiation Sciences, Umeå University, Umeå, Sweden

E-mail: lukas.fetty@meduniwien.ac.at

Keywords: synthetic CTs, radiation therapy, neural networks, GANs, cGAN architecture

Abstract

Recent developments in magnetic resonance (MR) to synthetic computed tomography (sCT) conversion have shown that treatment planning is possible without an initial planning CT. Promising conversion results have been demonstrated recently using conditional generative adversarial networks (cGANs). However, the performance is generally only tested on images from one MR scanner, which neglects the potential of neural networks to find general high-level abstract features. In this study, we explored the generalizability of the generator models, trained on a single field strength scanner, to data acquired with higher field strengths. T2-weighted 0.35T MRIs and CTs from 51 patients treated for prostate (40) and cervical cancer (11) were included. 25 of them were used to train four different generators (SE-ResNet, DenseNet, U-Net, and Embedded Net). Further, an ensemble model was created from the four network outputs. The models were validated on 16 patients from a 0.35T MR scanner. Further, the trained models were tested on the Gold Atlas dataset, containing T2-weighted MR scans of different field strengths; 1.5T(7) and 3T(12), and 10 patients from the 0.35T scanner. The sCTs were dosimetrically compared using clinical VMAT plans for all test patients. For the same scanner (0.35T), the results from the different models were comparable on the test set, with only minor differences in the mean absolute error (MAE) (35-51HU body). Similar results were obtained for conversions of 3T GE Signa and the 3T GE Discovery images (40-62HU MAE) for three of the models. However, larger differences were observed for the 1.5T images (48-65HU MAE). The overall best model was found to be the ensemble model. All dose differences were below 1%. This study shows that it is possible to generalize models trained on images of one scanner to other scanners and different field strengths. The best metric results were achieved by the combination of all networks.

1. Introduction

A milestone was recently reached in radiation therapy based solely on magnetic resonance images (MR-guided) with the commercialization of combined MR-linear accelerators (MR-LINAC) with simultaneous image acquisition and treatment (Raaymakers et al 2009, 2017, Mutic and Dempsey 2014). The electron densities are assigned regionally with bulk densities or by deformable image registration of the planning computed tomography (CT). For MR-only simulations, which require the conversion from MR images to synthetic CTs (sCT), commercial MR to sCT conversion tools are now available (Siversson et al 2015), (Köhler et al 2015). Those tools often rely on atlas- (Arabi et al 2016, Persson et al 2017, Guerreiro et al 2017) and regression-based approaches (Tyagi et al 2017), (Liu et al 2017) and have demonstrated
clinical feasibility. However, these methods have restrictions, such as the limitation of atlas pairs or the requirement of specific MR sequences.

The remarkable performance of deep learning methods in imaging applications, in particular by convolutional neural networks (CNNs) and generative adversarial networks (GANs), has increased the popularity of deep learning methods over a wide range of research fields. Recent review articles demonstrated that CNNs have been applied in a variety of radiation therapy applications and show at least comparable, if not better, results to existing methods (e.g. Meyer et al 2018, Sahiner et al 2019). The image-to-image translation task has been investigated by multiple groups for the purpose of MR to sCT conversion with GANs (Wolterink et al 2017, Emami et al 2018, Nie et al 2018, Maspero et al 2018) and CNNs (Nie et al 2016, Dinkla et al 2018, Xiang et al 2018). The application of deep learning has resulted in comparable outcomes regarding the mean absolute errors (MAEs) compared to atlas-based approaches. In the pelvic region, the performance of GANs has been slightly better than for CNNs (Nie et al 2018, Nie et al 2016). Further, treatment plan comparisons have demonstrated dose differences in the planning target volume (PTV) below 1% for the pelvic region (Dinkla et al 2018, Maspero et al 2018, Chen et al 2018).

A GAN consists of two models: a generator network that attempts to reproduce the input data distribution, and a discriminator network that attempts to classify samples as coming from the generator’s output distribution (e.g. sCTs) or from the input data distribution (e.g. original CT) (Goodfellow et al 2014). The two models are trained simultaneously attempting to beat each other. A conditional GAN (cGAN) is an extension of the GAN model in which both the generator and the discriminator are conditioned on some additional information, for instance a class label, or an input image modality such as in the case of MR-to-CT synthesis where the sCT image output would be conditioned on the MR image input. cGANs have a benefit in the image-to-image translation task due to the possibility to include a direct correlation of images (e.g. MR-CT) by conditioning on the MR image. One of the main advantages of cGANs is that the networks learn reasonable mappings even if the training dataset size is small, a notorious problem in medical image analysis (Isola et al 2017, Madani et al 2018).

Despite new cGAN methods being studied on a regular basis, it appears that there is only limited literature present investigating the effects of the different generator network architectures in the field of medical imaging. Additionally, the image-to-image translation is usually performed from one specific MR sequence to the planning CT, and the generalization has not been studied of the model to other sequences or different vendors and/or MR scanners operating at different field strengths. It is therefore beneficial to use an image dataset that includes multiple different MR scanner sequences and field strengths when evaluating the performance of an MR to sCT conversion tool. Such a dataset was created by Nyholm et al (2018), who collected MR and CT images from different centers with multiple observer delineations. This database can be used to define a general solution of the MR to sCT conversion problem by testing a model's ability to generalize to different MR scanners and imaging protocols, and even scanners with different field strengths.

In this study, we evaluated four different generator architectures in a cGAN for the task of converting T2-weighted MR images to sCT images. The networks were SE-ResNet, DenseNet, U-Net, and Embedded Net. Therefore, in this work we performed a systematic comparison of generator architectures trained on low field strength MR datasets in the attempt to find a general sCT conversion method that also works for other MR scanners with different field strengths, without the necessity to fine-tune the models to each new scanner type and field strength. We further evaluated an ensemble model where each ensemble output voxel was the median of the four network’s outputs. Ensemble models have been shown in numerous other applications to give better results than any of the individual models they were created from, and they usually do this by decreasing the variance and the bias of the model, compared to the individual models (Sagi and Rokach 2018). These benefits are usually explained as an approximation of the Bayes optimal model that would be obtained if all possible models from the hypothesis class could be included in the ensemble (Sagi and Rokach 2018). We investigated the generalization of these models to different MR scanners with different field strengths using the Gold Atlas dataset (Nyholm et al 2018).

2. Methods

The base architecture that was used for all the cGAN models in this work, was the pix2pix (Isola et al 2017) model. This architecture and its hyperparameters settings (including learning rate, discriminator depth, etc) were selected following Isola et al (2017) since it was demonstrated to have a low loss metric (on the translation task) for small datasets (~400 images). The goal of the work of Isola et al (2017) was to define a loss function without requiring further hand-engineering for new translation tasks. In the work of Isola et al different loss function and discriminator implementations were investigated to define values that can be applied to a broad range of different datasets; however, the authors used only two different generator architectures, namely the U-Net and an encoder-decoder network (Isola et al 2017). We did not change the
loss function or discriminator in any of the networks in this study, instead we built on the results of Isola et al (2017) with the aim of identifying the impact of the generator architecture on the conversion results.

The evaluation includes an existing architecture for the U-Net (Isola et al 2017) and adapted implementations of DenseNet (Huang et al 2017), SE-ResNet (Hu et al 2018), and Embedded Net (Xiang et al 2018) (see figure 1).

2.1. Network architecture
The general architecture was the pix2pix model (a cGAN extension) that uses a combination of a supervised L1 loss and a cGAN loss for the generator network. These loss functions were used as a similarity metric between generated and real images to determine the performance of the generator. It has been shown that this combination of loss functions achieves good results with minimal artifacts (Isola et al 2017). The cGAN loss function is defined as

\[ L_{cGAN}(G, D) = E_{x,y \sim \text{data}(x,y)} \left[ \log D(x, y) \right] + E_{x \sim \text{data}(x)} \left[ \log (1 - D(x, G(x))) \right], \]

where \( x \) is the input MR image and \( y \) is the input CT image, \( G \) is the generator network, \( D \) is the discriminator network that gives a prediction of the synthetic images while taking into account the true MR images. \( E_{x,y} \) and \( E_x \) are the expected values over the real data distributions \( P_{\text{data}}(x,y) \) and \( P_{\text{data}}(x) \), respectively. The aim of \( G(x) \) is to minimize the discriminator’s loss whereas \( D(x, y) \) aims to maximize the loss.

Further, the supervised L1 loss is defined as

\[ L_{L1}(G) = E_{x,y \sim \text{data}(x,y)} \left[ |||y - G(x)||_1 \right] \]

and measures the absolute deviations between the target image (i.e. CT) and the synthetically generated image (i.e. sCT). The final loss function is a weighted sum of the cGAN loss (equation 1) and the L1 loss (equation 2), i.e.

\[ G^* = \arg \min_G \max_D L_{cGAN}(G, D) + \lambda L_{L1}(G), \]

Figure 1. Visualisation of the networks’ generator architectures. The size of the boxes indicate the relative resolutions of the feature maps. The blue boxes represent convolutional layers and red boxes represent transposed convolutional layers, while green, yellow, and orange boxes represent the SE-ResNet, DenseNet, and Embedded blocks.
where \( \lambda < 0 \) controls the trade-off between the contribution of the unsupervised cGAN loss and the supervised L1 loss.

All implemented generators use a combination of convolution layers (C), instance normalization (I), and nonlinear activation functions (A) (ReLU, LeakyReLU, or PReLU). The instance normalization is used to reduce the internal covariate shift (Ioffee and Szegedy 2015). Variations in the parameters kernel size, stride, and padding are denoted as CIA(kx, sy, pz) for a combination of convolution, instance normalization, and activation where \( x, y, \) and \( z \) are the respective selected parameter values.

Four different generator architectures were evaluated, and they are briefly outlined in the following. The choice of these architectures was based on their recently reported outstanding performance in the literature, or frequent application in other studies as in the case of the U-Net. All of the models were trained in as 2D networks with 3 channel inputs which will be pointed out in more detail in section 2.5.

The U-Net architecture was first proposed in 2015 for the segmentation task (Ronneberger et al 2015). The U-Net has multiple benefits, for instance that previous activations are recycled through skip connections, which improves the gradient flow during backpropagation. The original U-Net had two convolution layers at each level, with \( 3 \times 3 \) filters in the convolutions, and finally a \( 2 \times 2 \) max pooling layer. In the pix2pix implementation (Isola et al 2017) one convolution layer is instead used with CIA(k4, s2, p1), i.e. the downsampling is achieved using a strided convolution instead of through max pooling.

The Embedded Net was implemented as described in Xiang et al (2018), with the exception of the loss function (cf equation 3). The general idea of this network was to reconstruct the CT from the learned features and embed this reconstruction into the feature maps at different points in the network by having multiple loss functions (see figure 2 in the supplementary material (https://stacks.iop.org/PMB/65/105004/mmedia)). The activation function of the reconstruction block was the hyperbolic tangent function (tanh) to ensure values between \(-1\) and \(1\). Five layers with a CIA(k3, s1, p1) structure and 128 feature maps were applied before the embedded blocks. As suggested by Xiang et al (2018), five embedded blocks were used. The advantage of the Embedded Net is a good performance with a relatively low number of network parameters.

The SE-ResNet (Squeeze-and-Excitation ResNet) (Hu et al 2018) is an extension of the ResNet architecture that includes an additional scaling factor for the feature maps that is learned by linear layers. The scaling factor squeezes all information into a single feature for each feature map by a global average pooling operation followed by a fully connected layer with a ReLU (rectified linear unit) activation. This technique describes a scaling of all feature maps by a factor plus a bias term that is defined by a skip connection.

DenseNet (Huang et al 2017) relies on the idea of the ResNet but instead of summing the information in the final step, this network concatenates the output feature maps from the layers with the input signal (see figure 2 in the supplementary material). In our implementation, the input is passed through two module blocks. The output is concatenated with the input features and this is further processed in the next module block. The size of the feature maps increases with the number of layers and the number of extracted features of the previous convolution. Finally, all feature maps are reduced to the initial number of features by a so-called transition block. In this implementation, nine dense blocks were used.

The ensemble model combines all four individually trained models. In theory, a combination of different networks would result in a more robust and statistical reliable model (Sagi and Rokach 2018). For instance, an ensemble model should avoid overfitting better because if one of the individual models overfits the data, the combination of several models will likely not be as affected. Similarly, any of the individual models might be trapped in a sub-optimal local minimum. These benefits have been observed several times in the literature (Sagi and Rokach 2018, Klement et al 2012, Huyhn et al 2016, Zhang et al 2018, Mak et al 2019). In this work the combination was done by computing the voxel-wise median. This should reduce the influence of outliers and reduce the variance of the predictions. Further, in this work, the network was applied to overlapping patches of the input image, which leads to a further increase in the number of proposals for each voxel. Having overlapping patches meant that each voxel was predicted four times, leading to a total of \(3 \times crops\) predictions for each voxel, and the ensemble prediction for that voxel would be the median of all those \(3 \times crops\) predictions.

The general schematic of the different generator architectures can be seen in figure 1, where it is shown that the SE-ResNet and the DenseNet have a similar down and up-sampling structure but differ in the block convolutions that are used. The convolution steps inside of the respective blocks are illustrated in figure 2 of the supplementary material. More information on the network implementations can be found in the supplementary material.

The discriminator was the same for all networks, i.e. the \(70 \times 70\) patch discriminator used in pix2pix (Isola et al 2017). The network includes three CIA(k4, s2, p1) blocks and one additional CIA(k4, s1, p1) block (Isola et al 2017). The final convolution reduces the number of feature maps to a single one, and outputs a \(30 \times 30\) prediction map of the discriminator.
Table 1. Acquisition parameters of the modalities for the different datasets.

| CT                        | train/eval/test set | test set 2 (3T GE Discovery) | test set 3 (1.5T Siemens) | test set 4 (3T GE Signa) |
|---------------------------|---------------------|------------------------------|---------------------------|--------------------------|
| Manufacturer              | Siemens             | Toshiba                      | Siemens                    | Siemens                   |
| Model                     | Somatom Definition  | Aquilion                     | Somatom Definition AS      | Emotion 6                 |
| Slice thickness (mm)       | 2                   | 3                            | 2                          | 2.5                       |
| Pixel size (mm²)           | 0.89 × 0.89         | 0.98 × 0.98                  | 1 × 1                      | 0.98 × 0.98               |
| Voltage (kV)               | 120                 | 120                          | 120                        | 120                       |
| Current (mA)               | 89–189              | 324–660                      | 104–294                    | 93–172                    |
| Exposure time (msec)       | 500                 | 500                          | 500                        | 500–700                   |
| Reconstruction diameter (mm)| 500                 | 500                          | 550–700                    | 500–600                   |
| T2w MR                    | Siemens             | GE                           | Siemens                    | GE                        |
| Model                     | Magnetom C!         | Discovery 750 w               | 1.5                        | Signa PET/MR              |
| Field strength (T)         | 0.35                | 3                            | 1.5                        | 3                         |
| Sequence type              | TSE                 | FRFSE                        | TSE                        | FRFSE                     |
| Slice thickness (mm)       | 5                   | 2.5                          | 2.5                        | 2.5                       |
| Pixel size (mm²)           | 1.25 × 1.25         | 0.875 × 0.875                | 0.875 × 1.1                | 0.875 × 0.875             |
| Bandwidth (Hz/pixel)       | 300                 | 390                          | 200                        | 390                       |
| TR (ms)                   | 3290                | 6000–6600                    | 12 000–16 000              | 9000                      |
| TE (ms)                   | 100                 | 97                           | 91–102                     | 65                        |
| Flip angle (°)             | 180                 | 111                          | 160                        | 111                       |
| Phase encoding direction   | COL                 | COL                          | ROW                       | COL                       |

2.2. Patient datasets

This study included 40 prostate and 11 cervix carcinoma patients. All patients were treated at the Department of Radiation Oncology, Medical University of Vienna/AKH, Vienna, Austria, between 2017 and 2018. The mean patient age was 75 years (56–82) for men and 51 years (30–78) for women. The tumor staging ranged from T1c to T2c (TNM Classification of Malignant Tumors) for the prostate and between 1b1 and 4b (Fédération Internationale de Gynécologie et d’Obstétrique FIGO staging) for the cervix cancer patients. All patients got volumetric modulated arc therapy (VMAT) treatment prescribed and followed a drinking protocol (900 ml of contrast solution 45 min before the scan). Fiducial markers were used for most male patients. All patients received a planning CT scan (SOMATOM Definition AS, Siemens Healthcare GmbH, Erlangen, Germany) followed by a 0.35T MR scan (Magnetom C!, Siemens Healthcare GmbH, Erlangen, Germany) with a flat table top (Witoszynskyj et al 2019) and respective immobilization aids. The CT protocol included a 120 kV spiral tomographic scan with a matrix size in the transversal plane of 512 × 512 pixels (0.89 × 0.89 mm²) and a slice thickness of 2 mm. The MR protocol included a T2-weighted 2D turbo spin echo (TSE) sequence with a matrix size of 320 × 320 pixels (1.25 × 1.25 mm²) and a slice thickness of 5 mm. The repetition and echo times were 2850 and 100 ms for each patient, respectively. The longitudinal field sizes varied for MR and CT from 15 to 30 cm dependent on the treatment protocol. None of the patients from the 0.35T machine had metal implants.

All networks were trained and evaluated with data from one scanner (0.35T MR). The Gold Atlas data set (Nyholm et al 2018) (from the Swedish Gentle Radiotherapy project) was used to test the generalization performance of the different networks. This dataset contains eight patients from a 3T GE Discovery 750 w (3T GE Discovery), seven patients from a 1.5T Siemens and four patients from a GE Signa positron emission tomography (PET)/MR scanner (3T GE Signa), all patients were scanned with a T2-weighted sequence. The CT image resolution varied from 0.89 × 0.89 to 1 × 1 mm, and the slice thickness varied from 2 to 3 mm. These data were used to test the final models and a more extended description of the data sets can be found in table 1.

The test dataset included one patient from the 1.5T scanner with a metal implant and one with contrast agent in the bladder. Further, one patient from the Gold Atlas dataset had aliasing artifacts (wrap-around artifacts) and was therefore excluded. The patient with the metal implant and the contrast agent was included the evaluation to demonstrate the efficiency of the model also for such patients.

2.3. Data preparation

The proportions of the 0.35T data were set to approximately 50% for training, 30% for validation and 20% for testing. All patient datasets were rigidly and further non-rigidly aligned using the Elastix registration...
software (Klein et al 2010), (Shamonin 2013). In the rigid and the non-rigid registration a multi-resolution pyramid scheme was used, with a resize factor of 4, 2, and 1. The adaptive stochastic gradient descent optimizer was used together with the normalized mutual information metric for the rigid and a combination of mutual information and transform bending energy penalty (ratio: 0.9999/0.0001) for the non-rigid registration. For the non-rigid registration, a predefined sampling region was used instead of a complete random sampling with 2048 sample points per iteration. MR data were corrected using the N4ITK bias field correction with the following parameter settings: convergence threshold 0.004, number of histogram bins 200, Wiener filter noise 0.01, down sampling 3. Additionally, a curvature anisotropic diffusion filter was applied to reduce noise (conductance 1, scaling update interval 1, time step 0.03, iterations 10). The intensity range was truncated from 0 to 2000 for MR and −600 to 1400 for CT and was further quantized to 8 bits. This conversion was performed since intensity values below −600 will be air and values above 1400 were identified only in a limited number of cases mainly in the femoral bones which is far from the dose planning region. The quantization corresponds to a difference of about 7.8 HU between gray values. A pre-evaluation showed minor effects from this, neglectable for the purpose of this study. The MR intensity truncation was selected near the maximum value where all voxels with an intensity value higher than 2000 was fat or bladder content. After the data preparation, the study contained a training set of 1506 images. The images were normalized to the range [−1, 1] before being fed into the networks. The Gold Atlas data was pre-processed similar to the training/validation/test set of the 0.35T machine. The CT volumes were already pre-registered by the research group. The data were resampled to the voxel size of the originally trained images (i.e. 359 × 359 pixels for both 3T MR scanners and 336 × 336 for the 1.5T Siemens scanner). Bias field correction and filtration was applied as described above. Normalization was done taking into account the different intensity ranges of the new data (analysis of the histograms).

For the data preparation and pre-/post-processing, workflows were defined and applied to the data using MICE Toolkit v1.0.7 (NONPI Medical AB, Umeå, Sweden).

2.4. Training
Training was performed on single transverse slices (2D), including the previous and next slices of the volume as separate channels in order to incorporate anatomical representations of the neighboring slices, and by that enhance the information content. For testing, the same technique was applied for each slice of the volume, which means that three predictions were made for each slice/crop in the longitudinal direction.

A hyperparameter search was conducted over the learning rate, number of epochs, and the trade-off parameter $\lambda$ in the ranges ($10^{-6}$, $10^{-3}$) on a logarithmic scale, (30, 100), and (10, 100), respectively. Bayesian optimization was used for the hyperparameter search with a Gaussian process regressor using limited memory BFGS as optimizer for an expected improvement acquisition function. As a reference, all networks were trained with the default settings of the pix2pix model, with $\lambda = 100$ (equation 3), 100 epochs, and a learning rate of $2 \times 10^{-4}$. All data were loaded with a resolution of $320 \times 320$ pixels. For data augmentation, the images were cropped randomly to $256 \times 256$ pixels. All weights were initialized from a Gaussian distribution with a mean of zero and a standard deviation of 0.02. Adam (Kingma and Ba 2015) was used for optimizing the weights, with momentum parameters of 0.5 and 0.999 for $\beta_1$ and $\beta_2$, respectively. The hyperparameter search was performed for 25 iterations and minimized the mean validation MAE over the last 10 epochs.

All training and testing was performed on a desktop computer with 32 GB DDR3 RAM and an AMD FX 4100 Quad-core CPU. The networks were implemented in PyTorch version 0.4 (Paszke et al 2017) and trained on a Nvidia TITAN Xp GPU.

2.5. Evaluation
To validate the performance of each of the architectures, three different settings were compared:

1. Image input size and number of crops: The images are given in full size without cropping (320 pixels), cropping to the training size (256 pixels) with 4 crops and 9 crops.
2. Hyperparameter: the default hyperparameters of the pix2pix model (Isola et al 2017) were compared to the best hyperparameters found by the hyper-parameter search.
3. Model parameters: The network weights of the best performing model of the last 10 epochs was compared to the average of the weights over the last 10 epochs (averaged weight model).

All pre-processed images were then fed into the network for prediction. The voxel-wise median was computed for the overlapping regions of the images of each slice to generate a single output image of the correct size. This led to 3, 12 and 27 prediction points for the no crop, 4 crops and 9 crops, respectively. The MAEs, the mean squared error (MSE), and the peak signal-noise ratio (PSNR) were computed on the final
predictions. The whole-body contour and the bone outline of the CT (meaning all voxels inside the body or bone outlines) were selected as regions of interest to exclude any artifacts introduced outside of the volume of interest. Bone was segmented by selecting all voxels above 250 HU. The results were further processed with a dilation and hole filling filter. These regions were used as masks within which the validation metrics were computed. Since voxel values below—600 HU were truncated, all voxels of the sCT with the value —600 HU were set to—1024 HU to prevent a dosimetric drawback and to agree with the CT calibration.

The conversion results were tested using the non-parametric Friedman Test followed by the Nemenyi post-hoc test (Demsar 2006). The statistical tests were carried out for the MAE of the whole body and the bone region comparing all pairs of the models with a significance level of 0.05.

The best models were tested on higher field strength MRs (i.e. on 1.5 and 3T). For this testing step, the Gold Atlas data set from the Swedish Gentle Radiotherapy project and 10 patients scanned in the 0.35T machine were used (Nyholm et al. 2018). The deviations in MR intensities were compensated for by truncating the maximum intensity to 400, 6500 and 4000 for 1.5, 3T GE and 3T Signa, respectively. The truncation values were selected according to a visual histogram analysis. The different MRs were compared for each network and scanner type within the body contour and the bone structures.

Treatment plans and resulting dose distributions (VMAT) were designed for the deformed planning CTs with one 360° arc with a beam energy of 10 MV for 39 fractions and 2 Gy per fraction (prescribed dose 78 Gy) in Raystation Research v6.99 (RaySearch, Stockholm, Sweden). The same settings were used to define a new radiation plan for the patients of our institute. All networks were used on the Gold Atlas dataset and the test set of the 0.35T scanner to generate sCTs and the designed plans were recomputed on the sCTs. The plans were designed to be clinical acceptable with the constraints given in this reference (Buschmann et al. 2016). Dose differences were recorded for the urinary bladder, rectum, left and right femoral heads, prostate, and the PTV. For the dose comparison the D2%, D50% and D98% were computed and translated to relative dose difference values by dividing the absolute difference by the prescribed dose.

The implant was spared during the planning to avoid the respective gantry positions. The patient with the aliasing artifact had to be excluded due to the incorrect representation of the external contour.

3. Results

3.1. Training and testing time
The DenseNet, SE-ResNet, Embedded Net, and U-Net networks were trained in 110 h, 22 h, 100 h, and 9 h, respectively. After the training, the MR to sCT conversion times were below 5 min, including the pre-processing steps. The conversion of a single image takes between 0.5 and 1 s on a single GPU, depending on the generator architecture. The total numbers of learned parameters for DenseNet, SE-ResNet, Embedded Net, and U-Net were $1.53 \times 10^7$, $1.15 \times 10^7$, $4.7 \times 10^7$, and $5.44 \times 10^7$, respectively. The discriminator had a total of $2.8 \times 10^8$ parameters in each network.

3.2. Validation and testing losses
The hyperparameter search resulted in other values than the default for each model. For the the learning rate (default $2 \times 10^{-4}$), the values ended up being $8 \times 10^{-4}$, $7.9 \times 10^{-4}$, $8.3 \times 10^{-4}$, and $9 \times 10^{-4}$ for the Embedded Net, SE-ResNet, DenseNet, and U-Net, respectively; the number of epochs (default 100) ended up being 98, 68, 40, and 30, respectively; and $\lambda$ (default 100) ended up being 100, 96, 10, and 81, respectively. However, when using these hyperparameters, the models ended up within one standard deviation of the models when using the default values in terms of MAE for whole body and bone. The averaged weight model gave the lowest mean MAE values for the whole body for all models except for the U-Net. Visual inspection of the U-Net results revealed that the conversions where not smooth but very noisy. Using different crops resulted in improved MAE values when higher numbers of crops were used (i.e. 9 crops of 256 pixels). The largest improvements were observed for the Embedded Net and U-Net. For consistency, we selected the default pix2pix hyperparameters for all models with the averaged weight model which were also further used to build the ensemble model. The results of the input image shape and the model parameter selection can be found in table 5 in the supplementary data.

For the test dataset, all networks performed mostly in the range of 33 to 51 HU MAE inside the whole-body contour for the 0.35T scanner images (see figure 2). The PSNR was almost equal for all networks with about 30.8 dB for the whole body and about 25 dB for the bone region. The ensemble model resulted in a MAE of about 41 HU. The statistical test performed using the Nemenyi post-hoc test showed significantly lower (p-value < 0.05) MAE between the ensemble model and the other networks, except the SE-ResNet for both the bone and body region.
Figure 2. MAE of the different scanners on the training set, test set, and the Gold Atlas dataset. The first row contains the test set errors of the 0.35T machine, the following rows contain the errors on the Gold Atlas dataset. In the first column the MAE in the whole-body region is depicted and in the second column the MAE in the bone regions. For each region and scanner, the different generator architectures are indicated by SE-ResNet (green), DenseNet (orange), U-Net (red), Embedded Net (black), and Ensemble model (blue).

Table 2. The table shows a comparison of the three investigated metrics (MAE, MSE, and PSNR). The tested networks (Embedded Net, U-Net SE-ResNet, DenseNet, and the ensemble model) and the different datasets are shown in the row.

| test dataset | models          | MAE (HU) | MSE (100 HU²) | PSNR (dB) |
|--------------|-----------------|----------|---------------|-----------|
|              |                 | body     | bone          | body      | bone      |                      |
| 0.35T Siemens| Ensemble model  | 41.2 ± 3.7| 121.2 ± 12.9  | 124.9 ± 29.6| 422.3 ± 84.1| 31.4 ± 1  | 26.1 ± 0.9 |
| Magnetom Ci! (test set 1) | SE-ResNet | 44.2 ± 4.3 | 127.0 ± 13.1  | 139.3 ± 32.9 | 454.6 ± 89.7 | 30.9 ± 1 | 25.8 ± 0.9 |
|              | DenseNet        | 44.8 ± 3.9 | 131.9 ± 11.2  | 143.6 ± 31.8 | 488.6 ± 82.1 | 30.8 ± 1 | 25.4 ± 0.8 |
|              | Embedded Net    | 44.8 ± 4.0 | 133.0 ± 13.3  | 142.6 ± 29.3 | 493.3 ± 94.9 | 30.8 ± 0.9 | 25.4 ± 0.9 |
|              | U-Net           | 45.1 ± 3.8 | 134.6 ± 13.6  | 142.8 ± 28.6 | 489.8 ± 93.6 | 30.8 ± 0.9 | 25.4 ± 0.9 |
| 1.5T Siemens| Ensemble model  | 52.0 ± 5.5 | 133.9 ± 22.3  | 201.4 ± 53.3 | 483.9 ± 226.9| 29.3 ± 1 | 25.8 ± 1.7 |
| (test set 2) | SE-ResNet       | 54.6 ± 5.8 | 144.1 ± 22.7  | 209.2 ± 54.3 | 542.0 ± 232.0| 29.2 ± 1 | 25.2 ± 1.6 |
|              | DenseNet        | 54.4 ± 5.5 | 137.6 ± 18.7  | 216.1 ± 54.6 | 493.9 ± 198.4| 29.0 ± 1 | 25.6 ± 1.5 |
|              | Embedded Net    | 59.8 ± 6.7 | 157.4 ± 25.6  | 241.5 ± 58.7 | 631.1 ± 259.8| 28.5 ± 1 | 24.6 ± 1.6 |
|              | U-Net           | 56.8 ± 5.5 | 145.2 ± 17.1  | 212.7 ± 54.6 | 528.2 ± 182.1| 29.1 ± 1 | 25.2 ± 1.3 |
| 3T GE        | Ensemble model  | 43.7 ± 6.2 | 122.9 ± 20.8  | 133.1 ± 35.0 | 355.9 ± 106.6| 31.1 ± 1 | 26.9 ± 1.3 |
| Discovery    | SE-ResNet       | 47.9 ± 6.8 | 132.8 ± 23.9  | 148.5 ± 36.0 | 404.8 ± 134.1| 30.7 ± 1 | 26.4 ± 1.4 |
| (test set 3) | DenseNet        | 47.7 ± 5.8 | 125.2 ± 14.4  | 148.7 ± 34.3 | 353.7 ± 68.4 | 30.6 ± 1 | 26.8 ± 0.9 |
|              | Embedded Net    | 59.1 ± 15.6 | 163.6 ± 28.1  | 225.1 ± 112.0 | 620.1 ± 181.9| 29.2 ± 2 | 24.5 ± 1.3 |
|              | U-Net           | 46.2 ± 5.9 | 134.0 ± 17.4  | 136.5 ± 32.7 | 390.5 ± 83.2 | 31.0 ± 1 | 26.0 ± 0.9 |
| 3T GE Sigma  | Ensemble model  | 48.2 ± 4.9 | 141.4 ± 35.5  | 146.2 ± 45.9 | 474.3 ± 226.4| 30.8 ± 1 | 25.9 ± 1.8 |
| (test set 4) | SE-ResNet       | 52.1 ± 5.8 | 150.6 ± 36.0  | 162.2 ± 56.3 | 519.8 ± 235.7| 30.4 ± 1 | 25.5 ± 1.7 |
|              | DenseNet        | 51.9 ± 4.7 | 142.3 ± 28.5  | 159.9 ± 41.0 | 461.1 ± 164.5| 30.3 ± 1 | 25.8 ± 1.4 |
|              | Embedded Net    | 63.9 ± 8.7 | 180.7 ± 41.0  | 245.0 ± 68.4 | 757.8 ± 346.9| 28.5 ± 1 | 23.8 ± 1.8 |
|              | U-Net           | 52.1 ± 5.1 | 153.5 ± 34.8  | 158.0 ± 46.5 | 524.0 ± 220.9| 30.4 ± 1 | 25.4 ± 1.6 |

3.3. Application of the pre-trained model for different field strengths

The results on the Gold Atlas data set are illustrated in table 2 and figure 2. Additional similarity metrics can be found in the supplementary material. The Embedded Net had larger MAE values (59–63 HU in average for the whole body) for all 3T images compared to the other models. The ensemble model resulted in the smallest MAE values in the bone region (134–141 HU on average) and in the whole body (43–52 HU on average).
Figure 3. Relative dose differences in the Gold Atlas dataset and the test set of the 0.35T scanner. The columns correspond to the scanners and the rows to the dose volume parameters. The color encodes the different models.

The Friedman test showed significant differences between the models for all datasets (p < 0.05). The pairwise testing with the Nemenyi test was significant for the comparison of the ensemble model to the Embedded Net and the U-Net for the 1.5T machine. Further the ensemble model showed significantly lower MAE values for the whole-body compared to all models for the 3T GE Discovery, except the U-Net, but only significantly lower values compared to the Embedded Net for the 3T GE Signa scanner (see figure 2).

A visual comparison of the best and worst examples, determined by the MAE for the different MR scanners, can be found in the supplementary material as well as a heatmap representation of all obtained p-values of the statistical evaluation.

3.4. Dosimetric evaluation
The dose differences are visualized in figure 3. The relative dose differences were below 1% for all scanners in the Gold Atlas dataset. The highest uncertainties were observed for the 0.35T MR scanner for the D2% where a general offset of about 0.5% was observed for the PTV. Overall the dose differences were below 1% for all models and all investigated structures except for the DenseNet which had errors for the bladder and the 1.5T machine up to nearly 1.5%. Similar uncertainties were also observed for the patient with the hip implant. Since dose plans were computed for all patients and models we collected a total of 140 data points for the doses.

4. Discussion
Multiple generator architectures were tested within a cGAN network to investigate the best conversion performance for the MR to sCT translation task in comparison to an ensemble model. These models were trained and validated on a dataset containing T2 weighted MR images from a 0.35T scanner. Subsequently, the trained models were applied to 1.5 and 3T images to evaluate model generalization. Note that two 3T machines of different vendors were investigated. The available literature provides no indication of the superiority of any generator architecture over the others for the task of sCT generation. In addition, it seems that the decision of using a certain network architecture was not, or only limitedly, outlined in the respective medical publications and only focuses on scanners available in the respective institution. In our study the
Figure 4. Comparison of the sCTs of the different generator models with the ground truth CT and the T2w MR of the 0.35T scanners test dataset. A difference map is given for the ensemble model and the ground truth CT.

0.35T MR images were acquired in-house, while the 1.5T and 3T MR images were obtained from the Gold Atlas dataset of the Swedish gentle radiotherapy project (Nyholm et al 2018).

Previous research has compared different network architectures for GANs and classifiers for non-medical applications (Zhang et al 2018, Mak et al 2019). They pointed out that better architectures do not necessarily learn features that can be directly applied to new datasets without fine-tuning. Kornblith et al (2019) demonstrated that the ResNet outperformed all networks when it was used as a fixed feature extractor for image classification. This behavior is to be expected in our study as well since in the area of image translation, the generator should learn features representing most of the patients’ anatomies. Lucic et al (2018) compared different GAN generator and discriminator loss function implementations and came to the conclusion that none of them outperformed the original system. Thus, Lucic et al recommend a systematic evaluation of GANs to define the most suitable network structure.

Another popular GAN approach is the CycleGAN, which learns features of unaligned data (Wolterink et al 2017), (Zhu et al 2017, Yang et al 2017). This is beneficial for sCT conversions since intermodal registration could be associated with substantial uncertainties (Ghose et al 2015). Further, the network can learn translations without the requirement of registered data if the conversion is performed in a cycle-consistent manner as described in detail in the paper of Zhu et al (2017). However, investigating CycleGAN was outside of the scope of this study.

We performed a hyperparameter search over the learning rate, the number of epochs, and the trade-off parameter \( \lambda \), which resulted in no significant improvements in comparison to the default settings of the pix2pix. However, other hyperparameters could potentially influence the performance of the models, such as, the training schedule for discriminator and generator, which are important in cGAN training, Adam momentum parameters and weight initialisation but also increase the necessity of thorough hyperparameter search. It is important to point out that additional hidden layers inside the discriminator have a direct influence on the total outcome due to the loss function implementation. This might be suboptimal for some generator configurations. These computing-intensive investigations, including architecture changes were, however, outside of the scope of this study.

The comparison of the 0.35T test set showed significant improvement of the MAE for the body as well as the bone for the ensemble model. No significance was only observed for the SE-ResNet architecture. However, all of the examined generators performed well for the trained scanner and showed no visual differences (see figure 4).

The results on the 0.35T MR scanner test sets were in good agreement with the literature. Xiang et al designed a CNN for brain and male pelvis MR to sCT conversion, resulting in a MAE of 42.5 ± 3.1 HU (Xiang et al 2018). More comparable architectures from Maspero et al and Nie et al using GANs for the translation task reported MAEs of 60 ± 6 HU and 39 ± 4.6 HU, respectively. All mentioned results are in a similar range to our generator implementations, where the MAE for SE-ResNet was 44.2 ± 4.3 HU, for DenseNet it was 44.8 ± 3.9 HU, for U-Net it was 45.1 ± 3.8 HU, for the Embedded net it was 44.8 ± 4.0 HU.
and for the ensemble model it was 41.2 ± 3.7 HU (see table 2). However, comparisons of MAE between different publications is challenging because it depends to great extend on the quality of the datasets, the registration accuracy, field of view, and image resolution. Maspero et al reported for example results for the whole body of 60 HU MAE in comparison to the MRIPlanner’s results of 36.5 HU MAE for male pelvic patient datasets where the percentage error of the dose inside of the PTV was below 1% for both cases and demonstrated clinical acceptance (Siversson et al 2015, Maspero et al 2018). Further, Maspero et al did not perform deformable registration which potentially could result in higher errors. This indicates that the quality of the MAE as a measure for sCT conversion might be limited when it comes to dosimetric evaluation. It is also worth mentioning that we investigated the direct relation between the MAE and the dose differences, which revealed no significant correlation (p > 0.05) between these two measures for the performed VMAT plans (see figure 5 in the supplementary materials). This may be influenced by the uniform dose deposition, but still leads to the assumption that the MAE is not a good metric to predict the accuracy of the dose calculation of sCTs. In this study, the validation focused on HU accuracy, however, geometric accuracy needs to be investigated as well to reach clinical implementation of sCTs.

The comparison of the conversion results of the 0.35T MR scanner to scanners with higher field strength demonstrated that the model generalization of the DenseNet and the SE-ResNet resulted in a better conversion in terms of the bone region (figure 2 and table 1). This is supported by the results of Kornblith et al (2019) who identified the ResNet architecture as the best fixed feature extractor. In their publication, multiple versions of VGG (Simonyan and Zisserman 2015), Inception (Szegedy et al 2015) and ResNet were included. However, SE-ResNet and DenseNet were not included. In the results presented here, the U-Net and the DenseNet performed best on the new scanner images where U-Net showed best results for the 3T GE Discovery systems (46.2 ± 5.9 HU) and DenseNet showed the best performance for the 3T GE Signa and the 1.5T images (51.9 ± 4.7 HU and 54.4 ± 5.5 HU MAE for the whole body). However, looking into the MAE of the bones the SE-ResNet and the DenseNet had lower mean values compared to the U-Net. The Embedded Net showed very good conversion capability for the trained field strength but had a large MAE (between 56 and 86 HU) for the whole body for 3T field strengths. Artifacts were also very likely to occur in the air region due to the differences in the noise level for the 0.35T MR and the 3T scanners. The misclassification of air can be explained by the training process where the network learns to ignore the noise. This had already been examined by Isola et al (2017). We performed tests where the noise outside the patient body was set to zero, which resulted in no significant difference in the MAE except for the Embedded Net. The MR image sequence type seems also to have an influence on the conversion. Both 3T machines used a fast recovery fast spin echo sequence while the 1.5T scanner used a TSE sequence like the 0.35T scanner. Additionally, the 3T devices were from a different vendor (GE) than the 1.5T and 0.35T units (Siemens). To what extent this influences the network conversion cannot easily be quantified, as it requires that different vendors provide very similar MR units. However, since MR does not acquire quantitative image information, it is important to investigate different sequence types or combinations of sequences to identify the best conversion performance for multiple MR scanners.

We have to address that the worst examples in the supplementary materials (figure 4) show patients with anatomical structures which were not well covered by the training dataset. Patients who had a low fat content or had a small pelvic were likely to fail during the conversion process. This is definitely a limitation due to the small dataset size even though the overall results are good despite this limitation.

The best overall performance over different MR scanners and field strengths was observed for the ensemble model, which achieved MAEs of 43.7 and 48.2 HU for the 3T scanners compared to 41.2 HU for the 0.35T scanner (i.e. the test set). The poorest performance was obtained for the 1.5T machine where the ensemble model resulted in an average MAE of 52.0 ± 5.5 HU but with min-max ranges of 45.9–61.7 HU. The high uncertainty can be explained by some misclassifications (e.g. muscle tissue to air). The MSE and PSNR showed a network performance similar to the MAE which demonstrated the best results for all datasets. The high uncertainty can be explained by some misclassifications (e.g. muscle tissue to air). The MSE and PSNR showed a network performance similar to the MAE which demonstrated the best results for all datasets. The principal advantages of such a model is the potential to increase the overall robustness. The ensemble model used in this study is only based on four individual models, but even so, a decreased MAE was observed, and this is likely a result that is general.

In this study a dosimetric comparison was performed between the sCT conversions of all models and the planning CT using the Gold Atlas dataset (i.e. 18 patients) and the test set of the 0.35T machine (i.e. 10 patients). All dose volume parameters of the ensemble model were below 1% and are comparable to previously performed studies (Siversson et al 2015, Maspero et al 2018, Dinkla et al 2018, Xiang et al 2018, Chen et al 2018, Edmund and Nyholm 2017, Han 2017).
not result in increased uncertainties in the dose planning when using the sCT instead of the CT. This could be explained by the cropping method which helps the network to increase HU value prediction of crops that do not include the hip implant (MAE of 53 HU and 178 HU for the body and bone for the ensemble model) and by sparing the hip implant during treatment planning. Further, high dose differences (D2%: PTV 3% and Rectum 5%) were observed for one patient measured with the 1.5T machine. The uncertainty was caused by contrast agent in the bladder as well as an increased amount of air in the rectum. After overwriting the CT intensities with water, the error dropped below 1%. This shows the general problem of changing patient anatomy during the treatment process. A systematic offset of the dose by 0.5% was observed for the 0.35T patients, however, this seemed to be due to inaccurate registrations of the CT and MR and not by conversion errors. For all other generator models the dose differences were similar low with the exception of the DenseNet for the 1.5T machine where we observed some higher errors up to 1.5%. This is expected since the conversion results looked visually comparable to the initially trained machine. Some conversions failed in terms of visual inspection (e.g. having the wrong anatomical representation), but the dosimetric comparison is still below 1% dose difference. This is a potential problem since a single visual inspection already showed that the sCT did not represent the anatomical structure of the MR. These results enhance our proposal of using neural networks for model generalisation and demonstrated that dedicated MR image features of one scanner can be transferred to other scanners with different field strength.

A thorough comparison of different GAN generator architectures has to the best of our knowledge not previously been performed in the literature. As shown in our study this choice has potential impact on the results of the image translation tasks, especially, when applying the network on data acquired with different settings (i.e. field strengths, scanner types, etc). The commonly used generator, the U-Net, did not demonstrate optimal results in our comparison. A common task for research groups is to use a U-Net-based architecture trained with in-house acquired data. The presented results are rarely compared to existing networks with good performance and further the quality of the data might vary to a great extent which influences the presented outcome more than the network design. Therefore, an open gold standard dataset, as used in the present study, is important in order to make models and the results comparable without the necessity of retraining the models on each dataset. Using in-house acquired data with multiple networks is only a proof of concept and cannot aim for general applicability of an MR-to-sCT conversion technique.

Finally, the cGAN architecture can be considered a useful model as it only needs a small sample size for the generator-discriminator cross-training. Also, using a network trained on 0.35T images for image conversion acquired at higher field strength showed promising results even though additional training might be required. Misclassification is probably the result of changing image parameters.

Future work should investigate the impact of the combination of the discriminator and generator losses for different network types. The implementation of a 3D cGAN has also the potential to increase the accuracy of the networks where not only three consecutive slices are considered but the full 3D volume of the patient. Further, the number of necessary training images should be investigated in a systematic comparison to define the impact of using GAN-based architectures in comparison to a classic CNN. The generalisation of the generator models could be further improved by including images from the Gold Atlas dataset during the training which could help to increase performance for models (i.e. Embedded Net).

5. Conclusion

In this study it was demonstrated that generator models pretrained on a 0.35T scanner can be used for scanners with different field strengths (1.5T to 3T) and from different vendors (Siemens and GE). The learned features of the 0.35T MR images showed comparable dosimetric results for the Gold Atlas dataset and the test set of the 0.35T machine to those in the literature.

Acknowledgments

The research was funded by the Austrian Science Fund (FWF) (project number: P30065-B27). We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

Disclosure of conflicts of interest

Tufve Nyholm and Tommy Löfstedt are co-owners of NONPI Medical AB, the developers of MICE Toolkit—the software for medical image analysis used in this study.
References

Arabi H, Koutsouvelis N, Rouzaud M, Miralbell R and Zaidi H 2016 Atlas-guided generation of pseudo-CT images for MRI-only and hybrid PET-MRI-guided radiotherapy treatment planning Phys. Med. Biol. 61 6531

Buschmann M, Seppenwoolde Y, Wiezerk T, Weiber K and Georg D 2016 Advanced optimization methods for whole pelvic and local prostate external beam therapy Phys. Med. Biol. 32 465–73

Chen S, Qin A, Zhou D and Yan D 2018 Technical Note: U-net-generated synthetic CT images for magnetic resonance imaging-only prostate intensity-modulated radiation therapy treatment planning Med. Phys. 45 5659–65

Demjar S 2008 Statistical comparisons of classifiers over multiple data sets J. Mach. Learn. Res. 7 1–30

Dinkla A M et al 2018 MR-only brain radiation therapy: dosimetric evaluation of synthetic CTs generated by a dilated convolutional neural network Int. J. Radiat. Oncol. Biol. Phys 102 801–12

Edmund J M and Nyholm T 2017 A review of substitute CT generation for MRI-only radiation therapy Radiat. Oncol. 12 28

Emami H, Dong M, Nejad-Davarani S P and Glade-Hurst C 2018 Generating synthetic CTs from magnetic resonance images using generative adversarial networks Med. Phys. 45 3627–36

Ghose S et al 2015 A review of segmentation and deformable registration methods applied to adaptive cervical cancer radiation therapy treatment planning Artif. Intell. Med. 64 75–87

Goodfellow I J, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A and Bengio Y 2014 Generative adversarial nets NIPS ’14: Proc. 27th Int. Conf. on Neural Information Processing Systems vol 2 ed Z Ghahramani et al (Cambridge, MA: MIT Press) pp 2672–80

Guerreiro F et al 2017 Evaluation of a multi-atlas CT synthesis approach for MRI-only radiotherapy treatment planning Phys. Medica. 35 7–17

Han X 2017 MR-based synthetic CT generation using a deep convolutional neural network method Med. Phys. 44 1408–19

Hu J, Shen L, Albani S, Sun G and Wu E 2018 Squeeze-and-excitation networks 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (Salt Lake City, UT, 18–23 June 2018) (Piscataway, NJ: IEEE) pp 7132–41

Huang G, Liu Z, Van Der Maaten L and Weinberger K Q 2017 Densely connected convolutional networks 2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (Honolulu, HI, 21–26 July 2017) (Piscataway, NJ: IEEE) pp 2261–9

Huynh T et al 2016 Estimating CT image from MRI data using structured random forest and auto-context model IEEE Trans. Med. Imaging 35 174–83

Joffe S and Szegedy C 2015 Batch normalization: accelerating deep network training by reducing internal covariate shift ICML–15: Proc. 32nd Int. Conf. on Machine Learning vol 37 ed F Blach and D Blei (MLJ) pp 448–56

Isola P, Zhu J Y, Zhou T and Efros A A 2017 Image-to-image translation with conditional adversarial networks 2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (Honolulu, HI, 21–26 July 2017) (Piscataway, NJ: IEEE) pp 5967–76

Kingma D P and Ba J L 2015 Adam: A method for stochastic optimization 3rd Int. Conf. on Learning Representations (ICLR 2015) (San Diego, CA, 7–9 May 2015) (arXiv:1412.6980)

Klein S, Staring M, Murphy K, Viergever M A and Pluim J 2010 elastix: A toolbox for intensity-based medical image registration IEEE Trans. Med. Imaging 29 196–205

Klement W, Wilk S, Michalowski W, Farion K J, Osmond M H and Verter V 2012 Predicting the need for CT imaging in children with minor head injury using an ensemble of Naive Bayes classifiers Artif. Intell. Med. 54 163–70

Köhler M, Vaara T, Van Grootel M, Hoogveen R, Kemppainen R and Renisch S 2015 MR-only simulation for radiotherapy planning: Philips MRCAT for prostate dose calculations using only MRI data Philips White Paper (Guildford: Philips)

Kornblith S, Shlens J and Le Q V 2019 Do better ImageNet models transfer better? 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR) (Long Beach, CA, 15–20 June 2019) (Piscataway, NJ: IEEE) pp 2560–66

Liu L, Jolly S, Cao Y, Vineberg K, Fessler J A and Balter J M 2017 Female pelvic synthetic CT generation based on joint intensity and shape analysis Phys. Med. Biol. 62 2935

Lucic M, Kurach K, Michalski M, Gelly S and Bousquet O 2018 Are GANs created equal? a large-scale study NIPS ’18: Proc. 32nd Int. Conf. on Neural Information Processing Systems ed S Bengio et al (New York: Curran Associates) pp 698–707

Madani A, Öng J R, Tiberai A and Mofrad M R K 2018 Deep echocardiography: data-efficient supervised and semi-supervised deep learning towards automated diagnosis of cardiac disease npj Digit. Med. 1 39

Mak R H et al 2019 Use of crowd innovation to develop an artificial intelligence-based solution for radiation therapy targeting JAMA Oncol. 5 654–61

Maspero M et al 2018 Dose estimation of fast synthetic-CT generation using a generative adversarial network for general pelvis MR-only radiotherapy Phys. Med. Biol. 63 185001

Meyer P, Noblet V, Mazzara C and Lallement E 2018 Survey on deep learning for radiotherapy Comput. Biol. Med. 98 126–46

Mutic S and Dempsy J F 2014 The ViewRay Treatment: magnetic resonance-guided and controlled radiotherapy Semin. Radiat. Oncol. 24 186–9

Nie D et al 2018 Medical image synthesis with deep convolutional adversarial networks IEEE Trans. Biomed. Eng. 65 2720–30

Nie D, Cao X, Gao Y, Wang L and Shen D 2016 Estimating CT image from MRI data using 3D fully convolutional networks Deep Learning and Data Labeling for Medical Applications: Proc. 1st Int. Workshop on Deep Learning in Medical Image Analysis (DLMIA 2016) and 2nd Int. Workshop on Large-ScaleAnnotation of Biomedical Data and Expert Label Synthesis (LABELS 2016) (Lecture Notes in Computer Science vol 10008) ed G Carneiro et al (Berlin: Springer) 170–8

Nyholm T et al 2018 MR and CT data with multiobserver delineations of organs in the pelvic area-Part of the Gold Atlas project Med. Phys. 45 1295–300

Paszke A et al 2017 Automatic differentiation in PyTorch NIPS 2017 Workshop Submission 31st. Conf. on Neural Information Processing Systems (Long Beach, CA, 4–9 December 2017)

Persson E et al 2017 MR-OPERA: A multicenter/multivendor validation of magnetic resonance imaging-only prostate treatment planning using synthetic computed tomography images Int. J. Radiat. Oncol. Biol. Phys 99 692–700
Raaymakers B W et al 2009 Integrating a 1.5 T MRI scanner with a 6 MV accelerator: proof of concept Phys. Med. Biol. 54 N229
Raaymakers B W et al 2017 First patients treated with a 1.5 T MRI-Linac: clinical proof of concept of a high-precision, high-field MRI guided radiotherapy treatment Phys. Med. Biol. 62 L41
Ronneberger O, Fischer P and Brox T 2015 U-Net: convolutional networks for biomedical image segmentation Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015 ed N Navab, J Hornegger, W M Wells and A F Frangi (Berlin: Springer) pp 234–41
Sagi O and Rokach L 2018 Ensemble learning: A survey WIREs Data Min. Knowl. Discov. 8 e1249
Sahiner B et al 2019 Deep learning in medical imaging and radiation therapy Am. Assoc. Phys. Med. 46 e1–36
Shamonin D 2013 Fast parallel image registration on CPU and GPU for diagnostic classification of Alzheimer’s disease Frontiers Neurinformed. 71–15
Simonyan K and Zisserman A 2015 Very deep convolutional networks for large-scale image recognition 3rd Int. Conf. on Learning Representations (ICLR 2015) (San Diego, CA, 7–9 May 2015) (arXiv:1409.1556v6)
Siversson C et al 2015 Technical note: MRI only prostate radiotherapy planning using the statistical decomposition algorithm Med. Phys. 42 6090–7
Spadea M F et al 2019 Deep convolution neural network (DCNN) multiplane approach to synthetic CT generation from MR images—application in brain proton therapy Int. J. Radiat. Oncol. Biol. Phys. 105 495–503
Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V and Rabinovich A 2015 Going deeper with convolutions 2015 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (Boston, MA, 7–12 June 2015) (Piscataway, NJ: IEEE) pp 1–9
Tyagi N et al 2017 Dosimetric and workflow evaluation of first commercial synthetic CT software for clinical use in pelvis Phys. Med. Biol. 62 2961
Witoszynskyj S et al 2019 Attenuation correction of a flat table top for radiation therapy in hybrid PET/MR using CT- and 68Ge/68Ga transmission scan-based μ-maps Phys. Medica. 65 76–83
Wolterink J M, Dinkla A M, Savenije M H F, Seevinck P R, van den Berg C A T and Ligum I 2017 Deep MR to CT synthesis using unpaired data Simulation and Synthesis in Medical Imaging: 2nd Int. Workshop (SASHIMI 2017) (Lecture Notes in Computer Science vol 10557) ed S A Tsaftaris, A Gooya, A F Frangi and J L Prince (Berlin: Springer) pp 14–23
Xiang L et al 2018 Deep embedding convolutional neural network for synthesizing CT image from T1-Weighted MR image Med. Image Anal. 47 31–44
Yang H et al 2017 Unpaired brain MR-to-CT synthesis using a structure-constrained CycleGAN Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th Int. Workshop (DLMIA 2018) and 8th Int. Workshop (ML-CDS 2018) (Lecture Notes in Computer Science vol 11045) ed D Stoyanov et al (Berlin: Springer) pp 174–82
Zhang J, Wu Q J, Xie T, Sheng Y, Yin -F-F and Ge Y 2018 An ensemble approach to knowledge-based intensity-modulated radiation therapy planning Frontiers Oncol. 8 1–9
Zhu J-Y, Park T, Isola P and Efros A A 2017 Unpaired image-to-image translation using cycle-consistent adversarial networks 2017 IEEE Int. Conf. on Computer Vision (ICCV) (Venice, Italy, 22–29 October 2017) (Piscataway, NJ: IEEE) pp 2242–51