

Deep MR to CT Synthesis using Unpaired Data

Jelmer M. Wolterink¹✉, Anna M. Dinkla², Mark H.F. Savenije²,
Peter R. Seevinck¹, Cornelis A.T. van den Berg², Ivana Išgum¹

¹ Image Sciences Institute, University Medical Center Utrecht, The Netherlands
j.m.wolterink@umcutrecht.nl

² Department of Radiotherapy, University Medical Center Utrecht, The Netherlands

Abstract. MR-only radiotherapy treatment planning requires accurate MR-to-CT synthesis. Current deep learning methods for MR-to-CT synthesis depend on pairwise aligned MR and CT training images of the same patient. However, misalignment between paired images could lead to errors in synthesized CT images. To overcome this, we propose to train a generative adversarial network (GAN) with unpaired MR and CT images. A GAN consisting of two synthesis convolutional neural networks (CNNs) and two discriminator CNNs was trained with cycle consistency to transform 2D brain MR image slices into 2D brain CT image slices and vice versa. Brain MR and CT images of 24 patients were analyzed. A quantitative evaluation showed that the model was able to synthesize CT images that closely approximate reference CT images, and was able to outperform a GAN model trained with paired MR and CT images.

Keywords: Deep learning, radiotherapy, treatment planning, CT synthesis, Generative Adversarial Networks

1 Introduction

Radiotherapy treatment planning requires a magnetic resonance (MR) volume for segmentation of tumor volume and organs at risk, as well as a spatially corresponding computed tomography (CT) volume for dose planning. Separate acquisition of these volumes is time-consuming, costly and a burden to the patient. Furthermore, voxel-wise spatial alignment between MR and CT images may be compromised, requiring accurate registration of MR and CT volumes. Hence, to circumvent separate CT acquisition, a range of methods have been proposed for MR-only radiotherapy treatment planning in which a substitute or synthetic CT image is derived from the available MR image [2].

Previously proposed methods have used convolutional neural networks (CNNs) for CT synthesis in the brain [4] and pelvic area [8]. These CNNs are trained by minimization of voxel-wise differences with respect to reference CT volumes that are rigidly aligned with the input MR images. However, slight voxel-wise misalignment of MR and CT images may lead to synthesis of blurred images. To address this, Nie et al. [9] proposed to combine the voxel-wise loss with an image-wise adversarial loss in a generative adversarial network (GAN) [3]. In this

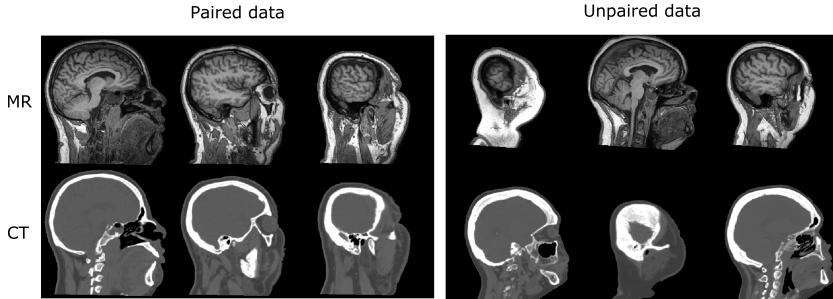


Fig. 1: *Left* When training with paired data, MR and CT slices that are simultaneously provided to the network correspond to the same patient at the same anatomical location. *Right* When training with unpaired data, MR and CT slices that are simultaneously provided to the network belong to different patients at different locations in the brain.

GAN, the synthesis CNN competes with a discriminator CNN that aims to distinguish synthetic images from real CT images. The discriminator CNN provides feedback to the synthesis CNN based on the overall quality of the synthesized CT images.

Although the GAN method by Nie et al. [9] addresses the issue of image misalignment by incorporating an image-wise loss, it still contains a voxel-wise loss component requiring a training set of paired MR and CT volumes. In practice, such a training set may be hard to obtain. Furthermore, given the scarcity of training data, it may be beneficial to utilize additional MR or CT training volumes from patients who were scanned for different purposes and who have not necessarily been imaged using both modalities. The use of unpaired MR and CT training data would relax many of the requirements of current deep learning-based CT synthesis systems (Fig. 1).

Recently, methods have been proposed to train image-to-image translation CNNs with unpaired natural images, namely DualGAN [11] and CycleGAN [12]. Like the methods proposed in [4,8,9], these CNNs translate an image from one domain to another domain. Unlike these methods, the loss function during training depends solely on the overall quality of the synthesized image as determined by an adversarial discriminator network. To prevent the synthesis CNN from generating images that look real but bear little similarity to the input image, cycle consistency is enforced. That is, an additional CNN is trained to translate the synthesized image back to the original domain and the difference between this reconstructed image and the original image is added as a regularization term during training.

Here, we use a CycleGAN model to synthesize brain CT images from brain MR images. We show that training with pairs of spatially aligned MR and CT images of the same patients is not necessary for deep learning-based CT synthesis.

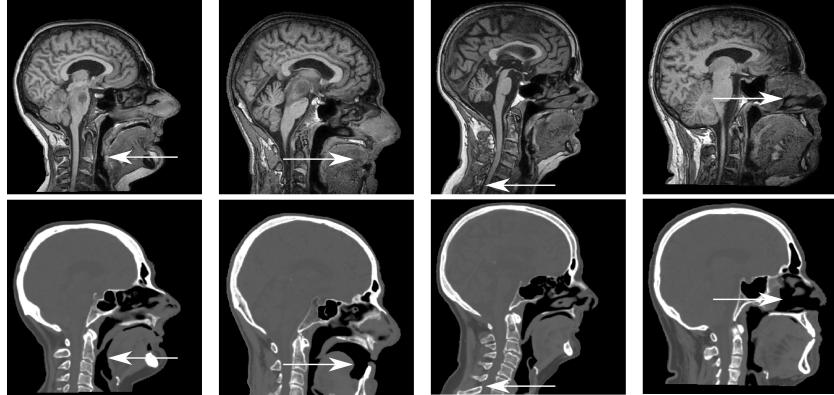


Fig. 2: Examples showing local misalignment between MR and CT images after rigid registration using mutual information. Although the skull is generally well-aligned, misalignments may occur in the throat, mouth, vertebrae, and nasal cavities.

2 Data

This study included brain MR and CT images of 24 patients that were scanned for radiotherapy treatment planning of brain tumors. MR and CT images were acquired on the same day in radiation treatment position using a thermoplastic mask for immobilization. Patients with heavy dental artefacts on CT and/or MR were excluded. T1 3D MR (repetition time 6.98 ms, echo time 3.14 ms, flip angle 8°) images were obtained with dual flex coils on a Philips Ingenia 1.5T MR scanner (Philips Healthcare, Best, The Netherlands). CT images were acquired helically on a Philips Brilliance Big Bore CT scanner (Philips Healthcare, Best, The Netherlands) with 120 kVp and 450 mAs. To allow voxel-wise comparison of synthetic and reference CT images, MR and CT images of the same patient were aligned using rigid registration based on mutual information following a clinical procedure. This registration did not correct for local misalignment (Fig. 2). CT images had a resolution of $1.00 \times 0.90 \times 0.90 \text{ mm}^3$ and were resampled to the same voxel size as the MR, namely $1.00 \times 0.87 \times 0.87 \text{ mm}^3$. Each volume had $183 \times 288 \times 288$ voxels. A head region mask excluding surrounding air was obtained in the CT image and propagated to the MR image.

3 Methods

The CycleGAN model proposed by Zhu et al. and used in this work contains a forward and a backward cycle (Fig. 3) [12].

The forward cycle consists of three separate CNNs. First, network Syn_{CT} is trained to translate an input MR image I_{MR} into a CT image. Second, network Syn_{MR} is trained to translate a synthesized CT image $Syn_{CT}(I_{MR})$ back into an

MR image. Third, network Dis_{CT} is trained to discriminate between synthesized $Syn_{CT}(I_{MR})$ and real CT images I_{CT} . Each of these three neural networks has a different goal. While Dis_{CT} aims to distinguish synthesized CT images from real CT images, network Syn_{CT} tries to prevent this by synthesizing images that cannot be distinguished from real CT images. These images should be translated back to the MR domain by network Syn_{MR} so that the original image is reconstructed from $Syn_{CT}(I_{MR})$ as accurately as possible.

To improve training stability, the backward cycle is also trained, translating CT images into MR images and back into CT images. For synthesis, this model uses the same CNNs Syn_{CT} and Syn_{MR} . In addition, it contains a discriminator network Dis_{MR} that aims to distinguish synthesized MR images from real MR images.

The adversarial goals of the synthesis and discriminator networks are reflected in their loss functions. The discriminator Dis_{CT} aims to predict the label 1 for real CT images and the label 0 for synthesized CT images. Hence, the discriminator Dis_{CT} tries to minimize

$$\mathcal{L}_{CT} = (1 - Dis_{CT}(I_{CT}))^2 + Dis_{CT}(Syn_{CT}(I_{MR}))^2 \quad (1)$$

for MR images I_{MR} and CT images I_{CT} . At the same time, synthesis network Syn_{CT} tries to maximize this loss by synthesizing images that cannot be distinguished from real CT images.

Similarly, the discriminator Dis_{MR} aims to predict the label 1 for real MR images and the label 0 for synthesized MR images. Hence, the loss function for MR synthesis that Dis_{MR} aims to minimize and Syn_{MR} aims to maximize is defined as

$$\mathcal{L}_{MR} = (1 - Dis_{MR}(I_{MR}))^2 + Dis_{MR}(Syn_{MR}(I_{CT}))^2 \quad (2)$$

To enforce bidirectional cycle consistency during training, additional loss terms are defined as the difference between original and reconstructed images,

$$\mathcal{L}_{Cycle} = \|Syn_{MR}(Syn_{CT}(I_{MR})) - I_{MR}\|_1 + \|Syn_{CT}(Syn_{MR}(I_{CT})) - I_{CT}\|_1. \quad (3)$$

During training, this term is weighted by a parameter λ and added to the loss functions for Syn_{CT} and Syn_{MR} .

3.1 CNN Architectures

The PyTorch implementation provided by the authors of [12] was used in all experiments³. This implementation performs voxel regression and image classification in 2D images. Here, experiments were performed using 2D sagittal image slices (Fig. 1). We provide a brief description of the synthesis and discriminator CNNs. Further implementation details are provided in [12].

³ <https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix>

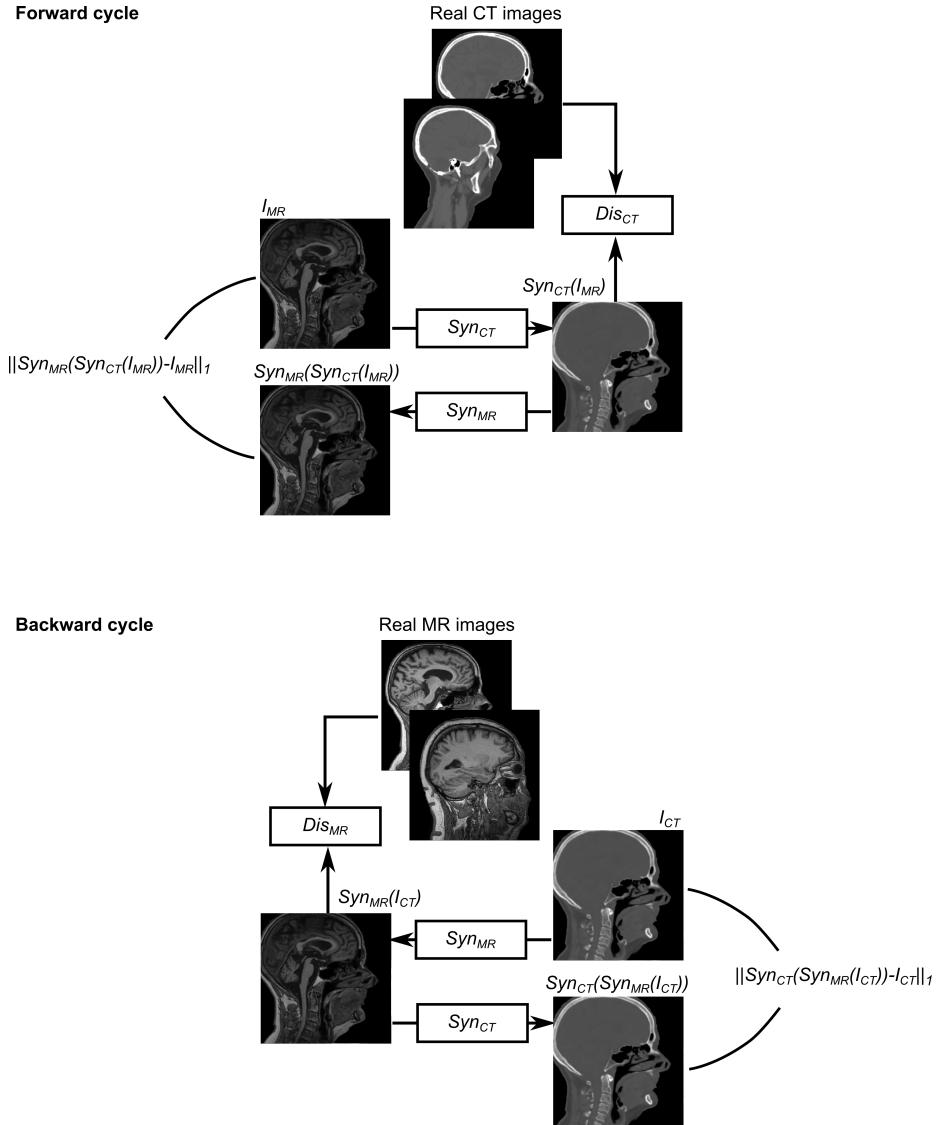


Fig. 3: The CycleGAN model consists of a forward cycle and a backward cycle. In the forward cycle, a synthesis network Syn_{CT} is trained to translate an input MR image I_{MR} into a CT image, network Syn_{MR} is trained to translate the resulting CT image back into an MR image that approximates the original MR image, and Dis_{CT} discriminates between real and synthesized CT images. In the backward cycle, Syn_{MR} synthesizes MR images from input CT images, Syn_{CT} reconstructs the input CT image from the synthesized image, and Dis_{MR} discriminates between real and synthesized MR images.

The network architectures of Syn_{CT} and Syn_{MR} are identical. They are 2D fully convolutional networks with two strided convolution layers, nine residual blocks and two fractionally strided convolution layers, based on the architecture proposed in [6] and used in [12]. Hence, the CNN takes input images of size 256×256 pixels and predicts output images of the same size.

Networks Dis_{CT} and Dis_{MR} also use the same architecture. This architecture does not provide one prediction for the full 256×256 pixel image, but instead uses a fully convolutional architecture to classify overlapping 70×70 image patches as real or fake [5]. This way, the CNN can better focus on high-frequency information that may distinguish real from synthesized images.

3.2 Evaluation

Real and synthesized CT images were compared using the mean absolute error

$$MAE = \frac{1}{N} \sum_{i=1}^N |I_{CT}(i) - Syn_{CT}(I_{MR}(i))|, \quad (4)$$

where i iterates over aligned voxels in the real and synthesized CT images. Note that this was based on the prior alignment of I_{MR} and I_{CT} . In addition, agreement was evaluated using the peak-signal-to-noise-ratio (PSNR) as proposed in [8,9] as

$$PSNR = 20 \log_{10} \frac{4095}{MSE}, \quad (5)$$

where MSE is the mean-squared error, i.e. $\frac{1}{N} \sum_{i=1}^N (I_{CT}(i) - Syn_{CT}(I_{MR}(i))^2$. The MAE and PSNR were computed within a head region mask determined in both the CT and MR that excludes any surrounding air.

4 Experiments and Results

The 24 data sets were separated into a training set containing MR and CT volumes of 18 patients and a separate test set containing MR and corresponding reference CT volumes of 6 patients.

Each MR or CT volume contained 183 sagittal 2D image slices. These were resampled to 256×256 pixel images with 256 grayscale values uniformly distributed in $[-600, 1400]$ HU for CT and $[0, 3500]$ for MR. This put image values in the same range as in [12], so that the default value of $\lambda = 10$ was used to weigh cycle consistency loss. To augment the number of training samples, each image was padded to 286×286 pixels and sub-images of 256×256 pixels were randomly cropped during training. The model was trained using Adam [7] for 100 epochs with a fixed learning rate of 0.0002, and 100 epochs in which the learning rate was linearly reduced to zero. Model training took 52 hours on a single NVIDIA Titan X GPU. MR to CT synthesis with a trained model took around 10 seconds.

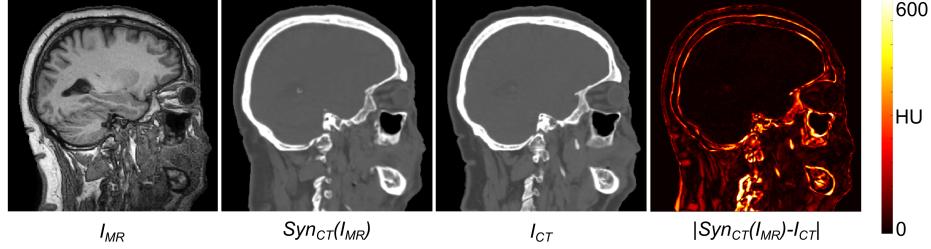


Fig. 4: *From left to right* Input MR image, synthesized CT image, reference real CT image, and absolute error between real and synthesized CT image.

Figure 4 shows an example MR input image, the synthesized CT image obtained by the model and the corresponding reference CT image. The model has learned to differentiate between different structures with similar intensity values in MR but not in CT, such as bone, ventricular fluid and air. The difference image shows the absolute error between the synthesized and real CT image. Differences are least pronounced in the soft brain tissue, and most in bone structures, such as the eye socket, the vertebrae and the jaw. This may be partly due to the reduced image quality in the neck area and misalignment between the MR image and the reference CT image. Table 1 shows a quantitative comparison between real CT and synthesized CT images in the test set. MAE and PSNR values show high consistency among the different test images.

To compare unpaired training with conventional paired training, an additional synthesis CNN with the same architecture as $Sync_{CT}$ was trained using paired MR and CT image slices. For this, we used the implementation of [5] which, like [9], combines voxel-wise loss with adversarial feedback from a discriminator network. This discriminator network had the same architecture as Dis_{CT} . A paired t-test on the results in Table 1 showed that agreement with the reference CT images was significantly lower ($p < 0.05$) for images obtained using this model than for images obtained using the unpaired model. Fig. 5 shows a visual comparison of results obtained with unpaired and paired training data. The image obtained with paired training data is more blurry and contains a high-intensity artifact in the neck.

During training, cycle consistency is explicitly imposed in both directions. Hence, an MR image that is translated to the CT domain should be successfully translated back to the MR domain. Fig. 6 shows an MR image, a synthesized CT image and the reconstructed MR image. The difference map shows that although there are errors with respect to the original image, these are very small and homogeneously distributed. Relative differences are largest at the contour of the head and in air, where intensity values are low. The reconstructed MR image is remarkably similar to the original MR image.

Table 1: Mean absolute error (MAE) values in HU and peak-signal-to-noise ratio (PSNR) between synthesized and real CT images when training with paired or unpaired data.

	MAE		PSNR	
	Unpaired	Paired	Unpaired	Paired
Patient 1	70.3	86.2	31.1	29.3
Patient 2	76.2	98.8	32.1	30.1
Patient 3	75.5	96.9	32.9	30.1
Patient 4	75.2	86.0	32.9	31.7
Patient 5	72.0	81.7	32.3	31.2
Patient 6	73.0	87.0	32.5	30.9
Average \pm SD	73.7 ± 2.3	89.4 ± 6.8	32.3 ± 0.7	30.6 ± 0.9

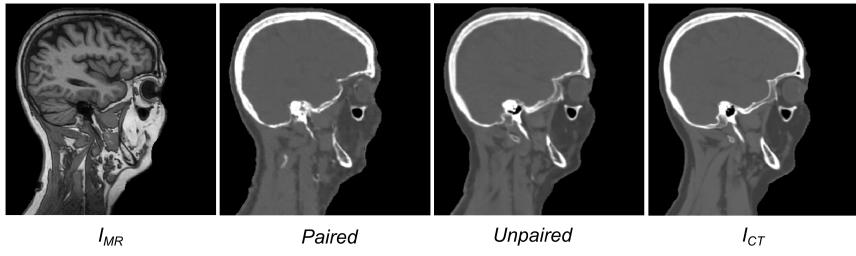


Fig. 5: *From left to right* Input MR image, synthesized CT image with paired training, synthesized CT image with unpaired training, reference real CT image.

5 Discussion and Conclusion

We have shown that a CNN can be trained to synthesize a CT image from an MR image using unpaired and unaligned MR and CT training images. In contrast to previous work, the model learns to synthesize realistically-looking images guided only by the performance of an adversarial discriminator network and the similarity of back-transformed output images to the original input image.

Quantitative evaluation using an independent test set of six images showed that the average correspondence between synthetic CT and reference CT images was 73.7 ± 2.3 HU (MAE) and 32.3 ± 0.7 (PSNR). In comparison, Nie et al. reported an MAE of 92.5 ± 13.9 HU and a PSNR of 27.6 ± 1.3 [9], and Han et al. reported an MAE of 84.8 ± 17.3 HU [4]. However, these studies used different data sets with different anatomical coverage, making a direct comparison infeasible. Furthermore, slight misalignments between reference MR and CT images, and thus between synthesized CT and reference CT, may have a large effect on quantitative evaluation. In future work, we will evaluate the accuracy of synthesized CT images in radiotherapy treatment dose planning.

Yi et al. showed that a model using cycle consistency for unpaired data can in some cases outperform a GAN-model on paired data [11]. Similarly, we found that in our test data sets, the model trained using unpaired data outperformed

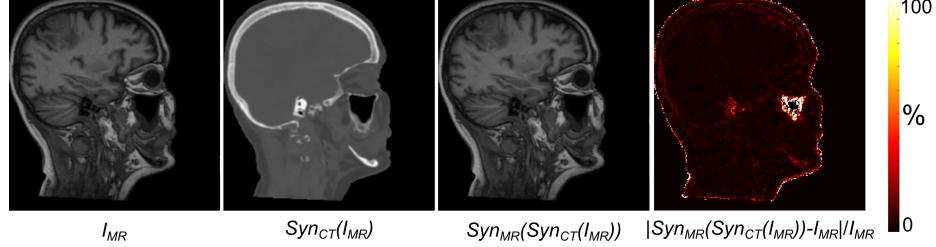


Fig. 6: *From left to right* Input MR image, synthesized CT image, reconstructed MR image, and relative error between the input and reconstructed MR image.

the model trained using paired data. Qualitative analysis showed that CT images obtained by the model trained with unpaired data looked more realistic, contained less artifacts and contained less blurring than those obtained by the model trained with paired data. This was reflected in the quantitative analysis. This could be due to misalignment between MR and CT images (Fig. 2), which is ignored when training with unpaired data.

The results indicate that image synthesis CNNs can be trained using unaligned data. This could have implications for MR-only radiotherapy treatment planning, but also for clinical applications where patients typically receive only one scan of a single anatomical region. In such scenarios, paired data is scarce, but there are many single acquisitions of different modalities. Possible applications are synthesis between MR images acquired at different field strengths [1], or between CT images acquired at different dose levels [10].

Although the CycleGAN implementation used in the current study was developed for natural images, synthesis was successfully performed in 2D medical images. In future work, we will investigate whether 3D information as present in MR and CT images can further improve performance. Nonetheless, the current results already showed that the synthesis network was able to efficiently translate structures with complex 3D appearance, such as vertebrae and bones.

The results in this study were obtained using a model that was trained with MR and CT images of the same patients. These images were rigidly registered to allow a voxel-wise comparison between synthesized CT and reference CT images. We do not expect this registration step to influence training, as training images were provided in a randomized unpaired way, making it unlikely that both an MR image and its registered corresponding CT image were simultaneously shown to the GAN. In addition, images were randomly cropped, which partially cancels the effects of rigid registration. Nevertheless, using images of the same patients in the MR set and the CT set may affect training. The synthesis networks could receive stronger feedback from the discriminator, which would occasionally see the corresponding reference image. In future work, we will extend the training set to investigate if we can similarly train the model with MR and CT images of *disjoint* patient sets.

References

1. Bahrami, K., Shi, F., Rekik, I., Shen, D.: Convolutional neural network for reconstruction of 7T-like images from 3T MRI using appearance and anatomical features. In: International Workshop on Large-Scale Annotation of Biomedical Data and Expert Label Synthesis. pp. 39–47. Springer (2016)
2. Edmund, J.M., Nyholm, T.: A review of substitute CT generation for MRI-only radiation therapy. *Radiation Oncology* 12(1), 28 (2017)
3. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems. pp. 2672–2680 (2014)
4. Han, X.: MR-based synthetic CT generation using a deep convolutional neural network method. *Medical Physics* 44(4), 1408–1419 (2017)
5. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. arXiv preprint arXiv:1611.07004 (2016)
6. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: European Conference on Computer Vision. pp. 694–711. Springer (2016)
7. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2015)
8. Nie, D., Cao, X., Gao, Y., Wang, L., Shen, D.: Estimating ct image from mri data using 3d fully convolutional networks. In: Carneiro, G., Mateus, D., Peter, L., Bradley, A., Tavares, J.M.R.S., Belagiannis, V., Papa, J.P., Nascimento, J.C., Loog, M., Lu, Z., Cardoso, J.S., Cornebise, J. (eds.) Deep Learning and Data Labeling for Medical Applications: First International Workshop, LABELS 2016, and Second International Workshop, DLMIA 2016, Held in Conjunction with MICCAI 2016, Athens, Greece, October 21, 2016, Proceedings. pp. 170–178. Springer International Publishing, Cham (2016)
9. Nie, D., Trullo, R., Petitjean, C., Ruan, S., Shen, D.: Medical image synthesis with context-aware generative adversarial networks. arXiv preprint arXiv:1612.05362 (2016)
10. Wolterink, J.M., Leiner, T., Viergever, M.A., Isgum, I.: Generative adversarial networks for noise reduction in low-dose CT. *IEEE Transactions on Medical Imaging* (2017)
11. Yi, Z., Zhang, H., Gong, P.T., et al.: Dualgan: Unsupervised dual learning for image-to-image translation. arXiv preprint arXiv:1704.02510 (2017)
12. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. arXiv preprint arXiv:1703.10593 (2017)