TPSDicyc: Improved Deformation Invariant Cross-domain Medical Image Synthesis

Chengjia Wang[1(✉)], Giorgos Papanastasiou[2], Sotirios Tsaftaris[3], Guang Yang[4], Calum Gray[2], David Newby[1], Gillian Macnaught[2], and Tom MacGillivray[2]

[1] BHF Centre for Cardiovascular Science, University of Edinburgh, Edinburgh EH16 4TJ, UK
chengjia.wang@ed.ac.uk
[2] Edinburgh Imaging Facility QMRI, University of Edinburgh, Edinburgh EH16 4TJ, UK
[3] Institute for Digital Communications, School of Engineering, University of Edinburgh, West Mains Rd, Edinburgh EH9 3FB, UK
[4] National Heart and Lung Institute, Imperial College London, London SW3 6LY, UK

Abstract. Cycle-consistent generative adversarial network (CycleGAN) has been widely used for cross-domain medical image systhesis tasks particularly due to its ability to deal with unpaired data. However, most CycleGAN-based synthesis methods can not achieve good alignment between the synthesized images and data from the source domain, even with additional image alignment losses. This is because the CycleGAN generator network can encode the relative deformations and noises associated to different domains. This can be detrimental for the downstream applications that rely on the synthesized images, such as generating pseudo-CT for PET-MR attenuation correction. In this paper, we present a deformation invariant model based on the deformation-invariant CycleGAN (DicycleGAN) architecture and the spatial transformation network (STN) using thin-plate-spline (TPS). The proposed method can be trained with unpaired and unaligned data, and generate synthesised images aligned with the source data. Robustness to the presence of relative deformations between data from the source and target domain has been evaluated through experiments on multi-sequence brain MR data and multi-modality abdominal CT and MR data. Experiment results demonstrated that our method can achieve better alignment between the source and target data while maintaining superior image quality of signal compared to several state-of-the-art CycleGAN-based methods.

1 Introduction

Cross-domain image synthesis is gaining popularity in a wide range of clinical applications to enable multi-modality synthesis without acquiring data from multiple modalities. However, the vast majority of cross-modality synthesis methods are solely evaluated on brain image data due to the low geometric variance. Otherwise, performance of the synthesis methods often rely on a registration-based

F. Knoll et al. (Eds.): MLMIR 2019, LNCS 11905, pp. 245–254, 2019.
https://doi.org/10.1007/978-3-030-33843-5_23

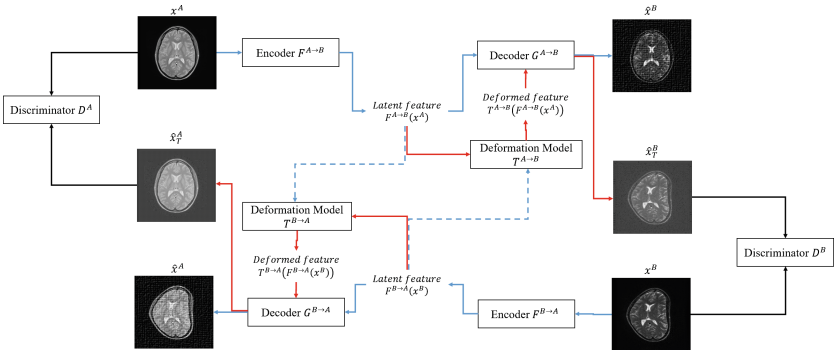


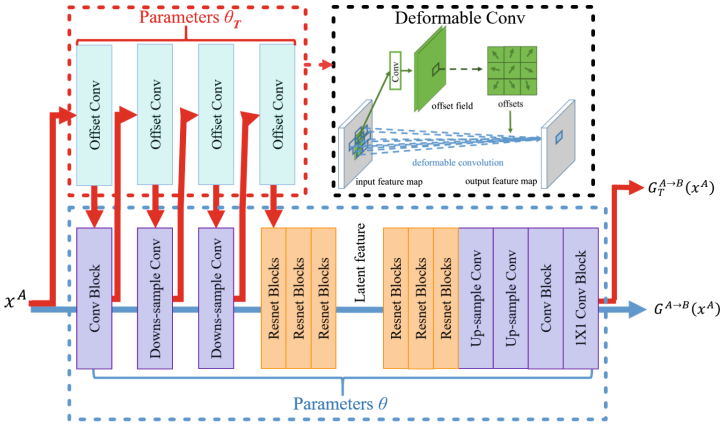
Fig. 1. Architecture of TPSDicyc

preprocessing step. Previous studies have shown that CycleGAN [1] achieves high synthesis quality on unpaired data, but it has also been observed that CycleGAN may reproduce the "domain-specific deformations" of the data [2,3]. A common strategy to address this issue is to leverage image similarity metrics in the CycleGAN loss function [4,5], but this introduces the trade-off between good quality of signal and good data alignment. The deformation-invariant CycleGAN (DicycleGAN) model [3] has been proposed recently achieved state-of-the-art synthesis performances with better alignment of data. This method uses two sets of parameters to encode translatable appearance features and relative spatial deformations between the training images using the deformable convolution (DC) operation [6]. However, DicycleGAN models a relatively consistent local deformation between the source and target data. This limits the generalizability of the model on multiple scanners, and requires the subjects in similar poses when being imaged. Otherwise the learning process can be unstable and slow to converge as shown in our experiments.

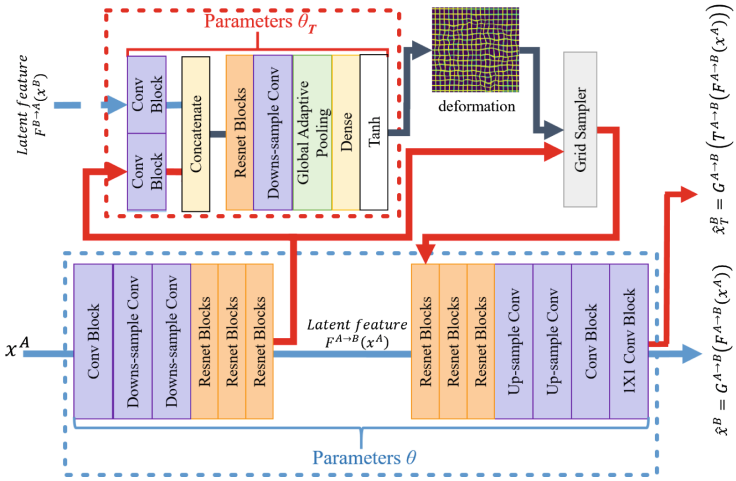
In this paper, we present an alternative framework of DicycleGAN based on thin-plate-spline (TPS) named as TPSDicyc. Compared to DicycleGAN, which models combines the "deformation" and "image translation" parameters into one network, TPSDicyc uses a separated spatial transformation network (STN) to learn the relative deformation between the source and target data. Figure 1 presents the TPSDicyc framework and its subnetworks. Figure 2 displays the architecture of TPSDicyc generator network. We evaluated the proposed method using both publicly available multi-sequence brain MR data and multi-modality abdominal data. Compared to the selected state-of-the-art baseline methods, TPSDicyc displayed better ability to handle disparate imaging domains and to generate synthesized images aligned with the source data.

2 Previous Works

CycleGAN was first applied to cross-domain medical image synthesis in [7] for co-synthesis of CT and MR brain data. Alignment between the synthesised data



(a) DicycleGAN generator



(b) TPSDicyc generator

Fig. 2. Structures and parameters of generators in DicycleGAN and our TPSDicyc: DicycleGAN uses deformable convolutional layers to model the relative deformation between the source and target data; in TPSDicyc the deformation is learned by a separated Thin-plate-spline based spatial transformation network. (Color figure online)

and the source data can be improved by regularizing the problem through multi-task training, for example, using segmentation masks [2], and by co-registration [7]. However, these models have an extra cost of manual annotations for segmentation or registration ground truths. A currently popular strategy is to integrate image similarity measures into the CycleGAN loss so that the geometric correspondences between data from different domains can be improved. For example, [5] introduces a structure-consistency loss based on the modality independent

neighborhood descriptor (MIND) [8]. It has been shown that this structure-constrained CycleGAN can be trained with unregistered multi-modal MR and CT brain data. A similar gradient-consistency loss, based on the normalised gradient cross correlation (GCC), is introduced in [4]. This method has been evaluated using unpaired but pre-registered, multi-modal MR and CT hip images. However, as discussed in [3], there is a conflict between the image similarity based losses and the CycleGAN discriminative loss. Because the synthesized data in which the "domain-specific doformations" are reproduced will lead to a lower adversarial loss (of the discriminator in GANs) but higher alignment loss. As a result, the synthesized data can not be well aligned to the source data and show a good quality of signal at the same time. DicycleGAN [3] uses DC parameters to decouple the translatable appearance features and the relative deformation between the source and target data, thus introduces a possible solution of the conflicts in the CycleGAN losses. However, DC layers can only learn relative consistent and local deformations.

3 Method

Assuming that we have n^A images $x^A \in \mathcal{X}^A$ from domain $\mathcal{X}^A$, and n^B images $x^B \in \mathcal{X}^B$ from domain $\mathcal{X}^B$, synthesis is performed to generate images of domain $\mathcal{X}^B$ using images from $\mathcal{X}^A$. To this end, we train a generator (which consists of an encoder $F^{A\to B}$, a decoder $G^{A\to B}$ and a STN $T^{A\to B}$) and a discriminator D^B in the min-max game of the GAN loss $\mathcal{L}_{GAN}\left(F^{A\to B}, T^{A\to B}, G^{A\to B}, D^B, \mathcal{X}^A, \mathcal{X}^B\right)$. We let $\mathcal{L}_{GAN}^{A\to B}$ denote this GAN loss for short and simple representation. Accordingly, $G^{B\to A}$, $F^{B\to A}$, $T^{B\to A}$, D^A, and the GAN loss $\mathcal{L}_{GAN}^{B\to A}$ are defined. A CycleGAN-based framework consists of two symmetric sets of generators act as mapping functions applied to a source domain, and two discriminators D^B and D^A to distinguish real and synthesized data for a target domain. The *cycle consistency* loss $\mathcal{L}_{cyc}^{A,B}$, is used to keep the cycle-consistency between the two sets of networks. This gives CycleGAN the ability to deal with unpaired data. Then the loss of the whole CycleGAN framework $\mathcal{L}_{CycleGAN}$ is $\mathcal{L}_{CycleGAN} = \mathcal{L}_{GAN}^{A\to B} + \mathcal{L}_{GAN}^{B\to A} + \lambda_{cyc}\mathcal{L}_{cyc}^{A,B}$. Presently proposed CycleGAN-based methods add an image alignment term $\mathcal{L}_{align}^{A,B}$ to $\mathcal{L}_{CycleGAN}$ which becomes $\mathcal{L}_{CycleGAN,align} = \mathcal{L}_{CycleGAN} + \lambda_{align}\mathcal{L}_{align}^{A,B} = \mathcal{L}_{GAN}^{A\to B} + \mathcal{L}_{GAN}^{B\to A} + \lambda_{cyc}\mathcal{L}_{cyc}^{A,B} + \lambda_{align}\mathcal{L}_{align}^{A,B}$, where λ_{align} is the weight used to balance $\mathcal{L}_{align}^{A,B}$ and $\mathcal{L}_{CycleGAN}$.

3.1 Architecture of Generator

As shown in Fig. 2, the generator network consists of an encoder, an decoder and a STN. The encoder maps the input data into a latent feature space, and the decoder generates the synthesized image based on the latent features. The relative deformation between the source and target domains are learned by a subset of parameters θ_T which are only used in the training process. In [3], θ_T is the trainable parameters in the DC layers. In this work, we introduced a new spatial transformation sub-network T for this purpose. As shown in Fig. 2, the transformation sub-network take latent features extracted

from the source and target data, and produces a displacement field of a keypoint grid. This displacement between keypoints are then applied to the source latent features using thin-plate-spline (TPS) interpolation. In this case, θ_T represents the parameters in this spatial transformation subnet, and θ the rest parameters in G. To generate synthesised images for domain $\mathcal{X}^B$, each input image x^A generates two output images through two separated forward passes: deformed output image $\hat{x}_T^B = G^{A\to B}\left(T^{A\to B}\left(F^{A\to B}\left(x^A\right)\right)\right)$ and undeformed image $\hat{x}^B = G^{A\to B}\left(F^{A\to B}\left(x^A\right)\right)$. $\hat{x}_T^B$ generated by passing the latent feature $F(x)$ through T (shown by the red arrows in Fig. 2) is expected to be identical to x^B. $\hat{x}^B$ is expected to be aligned with x^A.

3.2 Loss and Training

Similar to DicycleGAN, TPSDicyc loss functions include the traditional GAN loss, the cycle-consistency loss used in the original CycleGAN, an image alignment loss and an additional deformation invariant cycle consistency loss.

For the GAN loss $L_{GAN}^{A\to B}$, $F^{A\to B}$, $G^{A\to B}$ and $T^{A\to B}$ are trained to minimize $\left(D^B\left(\hat{x}_T^B\right)-1\right)^2$ and D^B is trained to minimize $\left(D^B(x^B)-1\right)^2 + D^B\left(\hat{x}_T^B\right)^2$. The same formulation is used to calculate $\mathcal{L}_{GAN}^{B\to A}$ defined on the "$B \to A$" networks. Note that the GAN loss is calculated based on the deformed outputs. As the undeformed outputs of generators are expected to be aligned with the input images, an image alignment loss based on normalized mutual information (NMI) is defined as:

$$\mathcal{L}_{align}^{A,B} = 2 - NMI\left(x^A, \hat{x}^B\right) - NMI\left(x^B, \hat{x}^A\right). \tag{1}$$

Essentially this image alignment loss can be adopted with any similarity measure suitable for image registration, such as normalized mutual information (NMI) [9], normalised GCC used in [4], or MIND in [5] and [8].

Secondly, the cycle-consistency loss plays a critical role for the outstanding performance of CycleGAN. In this work, both the undeformed and deformed version of synthesized data should be cycle-consistent to encode optimal representations. This results in two cycle-consistency losses. The undeformed cycle consistency loss is defined as:

$$\mathcal{L}_{cyc}^{A,B} = \|G^{B\to A}\left(F^{B\to A}\left(\hat{x}^B\right)\right) - x^A\|_1 + \|G^{A\to B}\left(F^{A\to B}\left(\hat{x}^A\right)\right) - x^B\|_1, \tag{2}$$

and the deformation-invariant cycle consistency loss is:

$$\begin{aligned}\mathcal{L}_{dicyc}^{A,B} =& \|G^{B\to A}\left(T^{B\to A}\left(F^{B\to A}\left(\hat{x}^B\right)\right)\right) - x^A\|_1 \\ &+ \|G^{A\to B}\left(T^{A\to B}\left(F^{A\to B}\left(\hat{x}^A\right)\right)\right) - x^B\|_1.\end{aligned} \tag{3}$$

The complete TPSDicyc loss is then defined as:

$$\mathcal{L}_{TPSDicyc} = \mathcal{L}_{GAN}^{A\to B} + \mathcal{L}_{GAN}^{B\to A} + \lambda_{align}\mathcal{L}_{align}^{A,B} + \lambda_{cyc}\mathcal{L}_{cyc}^{A,B} + \lambda_{dicyc}\mathcal{L}_{dicyc}^{A,B}. \tag{4}$$

In this work, we set $\lambda_{cyc} = \lambda_{dicyc} = 10$ and $\lambda_{align} = 0.9$. The models were trained with Adam optimizer [10] with a fixed learning rate of 0.0002 for the first 100 epochs, followed by 100 epochs with linearly decreasing learning rate. Here we apply a simple early stop strategy: in the first 100 epochs, when $\mathcal{L}_{TPSDicyc}$ stops decreasing for 10 epochs, the training will move to the learning rate decaying stage; similarly, this tolerance is set to 20 epochs in the second 100 epochs.

4 Experiments

IXI dataset: The Information eXtraction from Images (IXI) dataset[1] provides co-registered multi-sequence MR images collected from multiple sites. We used 66 pairs of proton density (PD-) and T2-weighted volumes for T2→PD synthesis experiment, each volume has 116 to 130 slices. We use 38 pairs for training and 28 pairs for evaluation of synthesis results. Our image generators take 2D axial-plane slices of the volumes as inputs. All volumes were resampled to a resolution of $1.8 \times 1.8 \times 1.8\,\mathrm{mm}^3$/voxel, then cropped to a size of 128×128 pixels. All the images are bias field corrected and normalized with their mean and standard deviation. We applied a simulated deformation to all T2-weighted images. Synthesis experiments were then performed between the undeformed PD-weighted data and the deformed T2-weighted data. When using deformed T2-weighted images, the ground truths of synthesized PD-weighted data were generated by applying the same nonlinear deformation to the source PD-weighted images.

Private Abdominal Data. We used a dataset containing 40 multi-modality abdominal T2*-weighted and CT images collected from 20 patients with abdominal aortic aneurysm (AAA). All images are resampled to a resolution of $1.56 \times 1.56 \times 5\,\mathrm{mm}^3$/voxel, and the axial-plane slices trimmed to 192×192 pixels. Because of the "domain-specific deformations", registration based ground truths as in the IXI dataset are not available. However, because several organs, such as aorta and spine, are relatively rigid compared to other surrounding soft tissues such as lower gastrointestinal tract organs, these objects can be affinely registered for evaluation of synthesis. For each volume in the MA^3RS dataset, the anatomy of the aorta were manually segmented for each volume (as described in [11]). The multi-modality data acquired from the same patient were affinely registered so that the segmented aorta in both data are well aligned. The manual registration and segmentation were performed by 4 clinical researchers. Signal of the synthesized images were evaluated within the segmentation of aorta.

4.1 Evaluation Metrics

To be consistent with the baseline methods, we use three metrics to evaluate performance on cross-domain image synthesis: mean squared error (MSE), peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) as typically

[1] http://brain-development.org/ixi-dataset/.

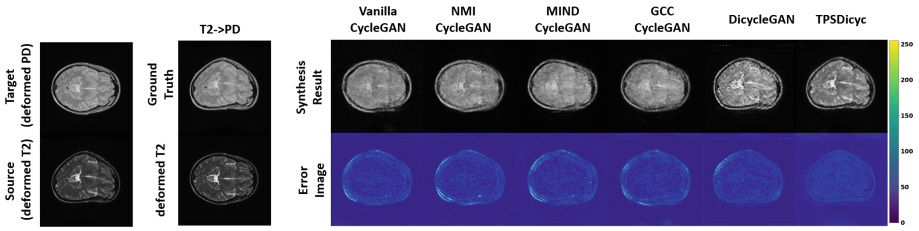


Fig. 3. Visualized PD→T2 synthesis results of the IXI dataset: an arbitrary deformation was applied to the T2 weighted images.

used by other CycleGAN based methods. Given a volume x^A and a target volume x^B, the MSE is computed as: $\frac{1}{N}\sum_1^N \left(x^B - \hat{x}^B\right)^2$, where N is number of voxels in the volume. PSNR is calculated as: $10\log_{10}\frac{\max_B^2}{MSE}$. SSIM is computed as: $\frac{(2\mu_A\mu_B+c_1)(2\delta_{AB}+c2)}{(\mu_A^2+\mu_B^2+c_1)(\delta_A^2+\delta+B^2+c2)}$, where μ and δ^2 are mean and variance of a volume, and δ_{AB} is the covariance between x^A and x^B. c_1 and c_2 are two variables to stabilize the division with weak denominator [12]. Larger PSNR and SSIM, or smaller MSE, indicate a better performance of a synthesis algorithm. To test the statistical significance of results, we perform paired t-test between the TPSDicyc and the DicycleGAN baseline. Differences between performances are considered to be statistically significant when the $p-$value is less than 0.05.

4.2 Results and Discussion

IXI Dataset. The quantitative results is shown in Table 1. Vanilla CycleGAN trained on paired and registered images (without simulated deformation) a theoretical upper-bound performance with PSNR > 24.3, SSIM > 0.817 and MSE ≤ 0.036. Trained with unpaired data sufferring from simulated deformation, the vanilla CycleGAN gave a lower-bound baseline of performances. With additive image alignment losses, GCC-CycleGAN [4] and MIND-CycleGAN [5] methods lead to tiny improvements in terms of PSNR. However, because these two models are still affected by the simulated "domain-specific deformation", their performances were still comparable to vanilla CycleGAN. In contrast, the proposed TPSDicyc model lead to results significantly closer to the upper-bound baseline.

Alignment between source and target data can be observed in the example shown in Fig. 3. It can be seen that the vanilla CycleGAN model exactly reproduced the simulated deformation. The GCC-CycleGAN and MIND-CycleGAN, although can reduce the misalignment effect, the synthesized and source data are still not well aligned. Furthermore, the synthesis results generated by the three CycleGAN-based models are blurry and showed visible artifacts. In contrast, our TPSDicyc model achieved best data alignment.

Abdominal data: Table 2 shows the quantitative assessments of the four compared models based on the same metrics used for the IXI data. The vanilla CycleGAN had slightly better performances compared the GCC- and MIND-CycleGAN models. Our method lead to over 20% performance gains in terms

Table 1. Synthesis results of IXI dataset using deformed T2 images.

	Method	MSE	PSNR	SSIM
	Cycle [7]	0.055 (0.22)	20.80 (2.87)	0.708 (0.19)
T2	GCC-Cycle [4]	0.054 (0.22)	21.04 (3.83)	0.719 (0.19)
↓	MIND-Cycle [5]	0.054 (0.21)	20.82 (2.61)	0.703 (0.19)
PD	DicycleGAN	0.045 (0.21)	22.52 (2.91)	0.790 (0.18)
	TPSDicyc	**0.044 (0.23)**	**22.72 (2.86)**	**0.796 (0.16)**
	Cycle (aligned)	0.037 (0.22)	24.77 (3.30)	0.856 (0.17)

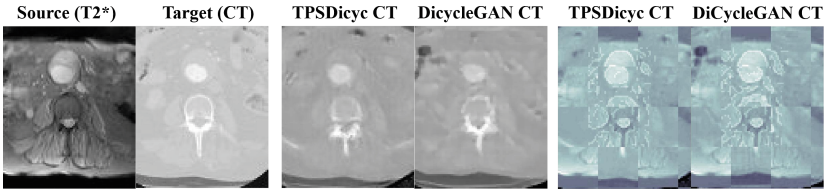


Fig. 4. Visualization of cross-modality synthesis results obtained with our MA^3RS dataset. A example data from both the CT and T2* domain are shown on the left. A checkerboard view combing the source and synthesized data is shown on the right. Alignment between the source and the synthesized data can then be assessed by looking at the anatomy of aorta and spine, as well as the lower contour of the patient body.

Table 2. T2*→CT synthesis results using private dataset.

T2* → CT			
Model	MSE	PSNR	SSIM
Cycle [7]	0.009 (0.004)	20.57 (2.12)	0.675 (0.06)
GCC-Cycle [4]	0.012 (0.006)	20.25 (2.35)	0.602 (0.08)
MIND-Cycle [5]	0.010 (0.004)	21.21 (2.04)	0.660 (0.07)
DicycleGAN	0.008 (0.004)	22.01 (2.40)	0.694 (0.07)
TPSDicyc	**0.008 (0.004)**	**22.29 (2.26)**	**0.706 (0.06)**

of MSE and SSIM, and also achieved better performance compared to Dicycle-GAN. Except for SSD, p-value of the paired t-test between DicycleGAN and our method are less than 0.05. Figure 4 provides a checkerboard visualization combining the source image and synthesized data generated by the DicycleGAN and our TPSDicyc. Objects such as spine and aorta in the source and target data can only be affinely registered independently. Both DicycleGAN and TPS-Dicyc model produce synthesized images where these objects are simultaneously aligned in the source and target data. DicycleGAN achieved better alignment of the outer contour of image subject while TPSDicyc show better alignment for spine and aorta.

5 Conclusion

In this paper, we propose the TPSDicyc model to address the issue of "domain-specific deformation". Different from the recently proposed DicycleGAN model, we integrate a TPS-based spatial transformation sub-network in the CycleGAN model and train the model with associated deformation-invariant cycle consistency loss and NMI-based alignment loss function. Compared to the DC layers in DicycleGAN, this new architecture allows to model global deformations. Our TPSDicyc method can achieve good alignment between the source and synthesized data, and outperformed the DicycleGAN, as well as state-of-the-art CycleGAN-based models in experiments performed on multi-sequence MR data and multi-modality abdominal data.

Acknowledgments. This work is funded by British Heart Fundation (no. RG/16/10/32375). S.A. Tsaftaris and G. Papanastasiou acknowledge support from the EPSRC Grant (EP/P022928/1). Support from NHS Lothian R&D, and Edinburgh Imaging and the Edinburgh Clinical Research Facility at the University of Edinburgh is gratefully acknowledged.

References

1. 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)
2. Chartsias, A., Joyce, T., Dharmakumar, R., Tsaftaris, S.A.: Adversarial image synthesis for unpaired multi-modal cardiac data. In: Tsaftaris, S.A., Gooya, A., Frangi, A.F., Prince, J.L. (eds.) SASHIMI 2017. LNCS, vol. 10557, pp. 3–13. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-68127-6_1
3. Wang, C., Macnaught, G., Papanastasiou, G., MacGillivray, T., Newby, D.: Unsupervised learning for cross-domain medical image synthesis using deformation invariant cycle consistency networks. In: Gooya, A., Goksel, O., Oguz, I., Burgos, N. (eds.) SASHIMI 2018. LNCS, vol. 11037, pp. 52–60. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00536-8_6
4. Hiasa, Y., et al.: Cross-modality image synthesis from unpaired data using cyclegan: effects of gradient consistency loss and training data size. arXiv preprint arXiv:1803.06629 (2018)
5. Yang, H., et al.: Unpaired brain MR-to-CT synthesis using a structure-constrained cyclegan. In: Stoyanov, D., et al. (eds.) DLMIA/ML-CDS -2018. LNCS, vol. 11045, pp. 174–182. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00889-5_20
6. Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y.: Deformable convolutional networks. CoRR, abs/1703.06211, vol. 1, no. 2 (2017). 3
7. Wolterink, J.M., Dinkla, A.M., Savenije, M.H.F., Seevinck, P.R., van den Berg, C.A.T., Išgum, I.: Deep MR to CT synthesis using unpaired data. In: Tsaftaris, S.A., Gooya, A., Frangi, A.F., Prince, J.L. (eds.) SASHIMI 2017. LNCS, vol. 10557, pp. 14–23. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-68127-6_2
8. Heinrich, M.P., et al.: Mind: modality independent neighbourhood descriptor for multi-modal deformable registration. Med. Image Anal. **16**(7), 1423–1435 (2012)

9. Vinh, N.X., Epps, J., Bailey, J.: Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance. J. Mach. Learn. Res. **11**(Oct), 2837–2854 (2010)
10. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
11. Papanastasiou, G., et al.: Multidimensional assessments of abdominal aortic aneurysms by magnetic resonance against ultrasound diameter measurements. In: Valdés Hernández, M., González-Castro, V. (eds.) MIUA 2017. CCIS, vol. 723, pp. 133–143. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-60964-5_12
12. Hore, A., Ziou, D.: Image quality metrics: PSNR vs. SSIM. In: 2010 20th international conference on Pattern recognition (ICPR), pp. 2366–2369. IEEE (2010)