Cross-Modality Image Synthesis from Unpaired Data Using CycleGAN

Effects of Gradient Consistency Loss and Training Data Size

Yuta Hiasa[1](✉), Yoshito Otake[1], Masaki Takao[2], Takumi Matsuoka[1], Kazuma Takashima[2], Aaron Carass[3], Jerry L. Prince[3], Nobuhiko Sugano[2], and Yoshinobu Sato[1]

[1] Graduate School of Science and Technology,
Nara Institute of Science and Technology,
8916-5, Takayamacho, Ikomashi, Nara 630-0192, Japan
hiasa.yuta.ht7@is.naist.jp

[2] Graduate School of Medicine, Osaka University, Suita, Japan

[3] Department of Electrical and Computer Engineering,
Johns Hopkins University, Baltimore, USA

Abstract. CT is commonly used in orthopedic procedures. MRI is used along with CT to identify muscle structures and diagnose osteonecrosis due to its superior soft tissue contrast. However, MRI has poor contrast for bone structures. Clearly, it would be helpful if a corresponding CT were available, as bone boundaries are more clearly seen and CT has a standardized (i.e., Hounsfield) unit. Therefore, we aim at MR-to-CT synthesis. While the CycleGAN was successfully applied to unpaired CT and MR images of the head, these images do not have as much variation of intensity pairs as do images in the pelvic region due to the presence of joints and muscles. In this paper, we extended the CycleGAN approach by adding the gradient consistency loss to improve the accuracy at the boundaries. We conducted two experiments. To evaluate image synthesis, we investigated dependency of image synthesis accuracy on (1) the number of training data and (2) incorporation of the gradient consistency loss. To demonstrate the applicability of our method, we also investigated segmentation accuracy on synthesized images.

Keywords: Image synthesis · CycleGAN · Musculoskeletal image MR · CT · Segmentation

1 Introduction

Computed tomography (CT) is commonly used in orthopedic procedures. Magnetic resonance imaging (MRI) is used along with CT to identify muscle structures and diagnose osteonecrosis due to its superior soft tissue contrast [1]. However, MRI has poor contrast for bone structures. It would be helpful if a corresponding CT were available, as bone boundaries are more clearly seen and CT

A. Gooya et al. (Eds.): SASHIMI 2018, LNCS 11037, pp. 31–41, 2018.
https://doi.org/10.1007/978-3-030-00536-8_4

has standardized (i.e., Hounsfield) units. Considering radiation exposure in CT, it is preferable if we can delineate boundaries of both muscle and bones in MRI. Therefore, we aim at MR-to-CT synthesis.

Image synthesis has been extensively studied using the patch-based learning [2] as well as deep learning, specifically, convolutional neural networks (CNN) [3] and generative adversarial networks (GAN) [4]. The conventional approaches required the paired training data, i.e., images of the same patient from multiple modalities that are registered, which limited the application. A method recently proposed by Zhu et al. [5], called CycleGAN, utilizes the unpaired training data by appreciating the cycle consistency loss function. While CycleGAN has already applied to MR-to-CT synthesis [6], all these previous approaches in medical image application targeted CT and MRI of the head in which the scan protocol (i.e., field-of-view (FOV) and the head orientation within the FOV) is relatively consistent resulting in a small variation in the two image distributions even without registration, thus a small number of training data set (20 to 30) allowed a reasonable accuracy. On the other hand, our target anatomy, the hip region, has larger variation in the anatomy as well as their pose (i.e., joint angle change and deformation of muscles).

Applications of image synthesis include segmentation. Some previous studies aimed at segmentation of musculoskeletal structures in MRI [7,8], but the issues in these studies were the requirement for multiple sequences and devices. Another challenge in segmentation of MRI is that there is no standard unit as in CT. Therefore, manually traced label data are necessary for training of each sequence and each imaging device. Thus, MR-to-CT synthesis realizes modality independent segmentation [9].

In this study, we extend the CycleGAN approach by adding the gradient consistency (GC) loss to encourage edge alignment between images in the two domains and using an order-of-magnitude larger training data set (302 MR and 613 CT volumes) in order to overcome the larger variation and improve the accuracy at the boundaries. We investigated dependency of image synthesis accuracy on 1) the number of training data and 2) incorporation of the GC loss. To demonstrate the applicability of our method, we also investigated a segmentation accuracy on synthesized images.

2 Method

2.1 Materials

The datasets we used in this study are MRI dataset consisting of 302 unlabeled volumes and CT dataset consisting of 613 unlabeled, and 20 labeled volumes which are associated with manual segmentation labels of 19 muscles around hip and thigh, pelvis, femur and sacrum bones. Patients with metallic artifact due to implant in the volume were excluded. As an evaluation dataset, we also used other three sets of paired MR and CT volumes, and 10 MR volumes associated with manual segmentation labels of gluteus medius and minimus muscles, pelvis and femur bones, as a ground truth. MR volumes were scanned in the coronal

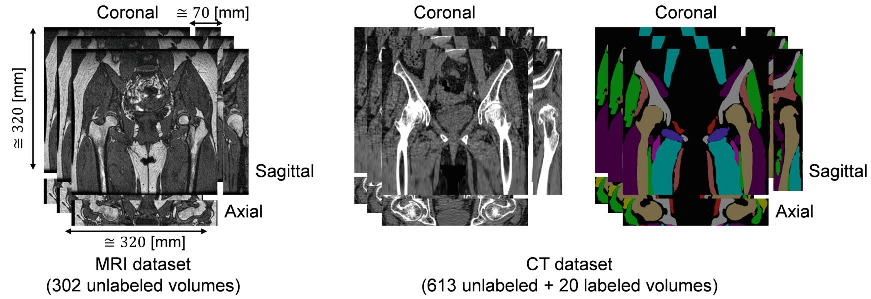


Fig. 1. Training datasets used in this study. MRI dataset consists of 302 unlabeled volumes and CT dataset consists of 613 unlabeled and 20 labeled volumes. N4ITK intensity inhomogeneity correction [10] was applied to all MRI volumes. Two datasets have similar field-of-view, although these are not registered.

plane for diagnosis of osteonecrosis by a 1.0T MR imaging system. The T1-weighted volumes were obtained by 3D spoiled gradient recalled echo sequence (SPGR) with a repetition time (TR) of 7.9 ms, echo time (TE) of 3.08 ms, and flip angle of 30. The field of view was 320 mm, and the matrix size was 256 × 256. The slab thickness was 76 mm, and the slice thickness was 2 mm without an inter-slice gap. CT volumes were scanned in the axial plane for diagnosis of the patients subjected to total hip arthroplasty (THA) surgery. The field of view was 360×360 mm and the matrix size was 512×512. The slice thickness was 2.0 mm for the region including pelvis and proximal femur, 6.0 mm for the femoral shaft region, and 1.0 mm for the distal femur region. In this study, the CT volumes were cropped and resliced so that the FOV resembles that of MRI volumes, as shown in Fig. 1, and then resized to 256 × 256.

2.2 Image Synthesis Using CycleGAN with Gradient-Consistency Loss

The underlying algorithm of the proposed MR-to-CT synthesis follows that of Zhu et al. [5] which allows to translate an image from CT domain to MR domain without pairwise aligned CT and MR training images of the same patient. The workflow of the proposed method is shown in Fig. 2. The networks G_{CT} and G_{MR} are generators to translate real MR and CT images to synthesized CT and MR images, respectivery. The networks D_{CT} and D_{MR} are discriminators to distinguish between real and synthesized images. While discriminators try to distinguish synthesized images by maximizing adversarial losses $\mathcal{L}_{CT}$ and $\mathcal{L}_{MR}$, defined as

$$\mathcal{L}_{CT} = \textstyle\sum_{x \in I_{CT}} \log D_{CT}(x) + \sum_{y \in I_{MR}} \log(1 - D_{CT}(G_{CT}(y))), \tag{1}$$

$$\mathcal{L}_{MR} = \textstyle\sum_{y \in I_{MR}} \log D_{MR}(y) + \sum_{x \in I_{CT}} \log(1 - D_{MR}(G_{MR}(x))), \tag{2}$$

generators try to synthesize images which is indistinguishable from the target domain by minimizing these losses. Where x and y are images from domains

I_{CT} and I_{MR}. However, networks with large capacity have potential to converge to the one that translate the same set of images from source domain to any random permutation of images in the target domain. Thus, adversarial losses alone cannot guarantee that the learned generator can translate an individual input to a desired corresponding output. Therefore, the loss function is regularized by cycle consistency, which is defined by the difference between real and reconstructed image, which is the inverse mapping of the synthesized image [5]. The cycle consistency loss $\mathcal{L}_{Cycle}$ is defined as

$$\mathcal{L}_{Cycle} = \textstyle\sum_{x \in I_{CT}} |G_{CT}(G_{MR}(x)) - x| + \sum_{y \in I_{MR}} |G_{MR}(G_{CT}(y)) - y| \quad (3)$$

We extended the CycleGAN approach by explicitly adding the gradient consistency loss between real and synthesized images to improve the accuracy at the boundaries. The gradient correlation (GC) [11] has been used as a similarity metric in the medical image registration, which is defined by the normalized cross correlation between two images. Given gradients in horizontal and vertical directions of these two images, A and B, GC is defined as

$$GC(A, B) = \frac{1}{2}\{NCC(\nabla_x A, \nabla_x B) + NCC(\nabla_y A, \nabla_y B)\} \quad (4)$$

$$\text{where,} \ \ NCC(A, B) = \frac{\sum_{(i,j)}(A - \bar{A})(B - \bar{B})}{\sqrt{\sum_{(i,j)}(A - \bar{A})^2}\sqrt{\sum_{(i,j)}(B - \bar{B})^2}}$$

and ∇_x and ∇_y are the gradient operator of each direction, $\bar{A}$ is the mean value of A. We formulate the gradient-consistency loss $\mathcal{L}_{GC}$ as

$$\mathcal{L}_{GC} = \frac{1}{2}\{\sum_{x \in I_{CT}} (1 - GC(x, G_{MR}(x))) + \sum_{y \in I_{MR}} (1 - GC(y, G_{CT}(y)))\} \quad (5)$$

Finally, our objective function is defined as:

$$\mathcal{L}_{total} = \mathcal{L}_{CT} + \mathcal{L}_{MR} + \lambda_{Cycle}\mathcal{L}_{Cycle} + \lambda_{GC}\mathcal{L}_{GC} \quad (6)$$

where λ_{Cycle} and λ_{GC} are weights to balance each loss. Then, we solve:

$$\hat{G}_{MR}, \hat{G}_{CT} = \arg \min_{G_{CT}, G_{MR}} \max_{D_{CT}, D_{MR}} \mathcal{L}_{total} \quad (7)$$

In this paper, we used 2D CNN with 9 residual blocks for generator, similar to the one proposed in [12]. For discriminators, we used 70 × 70 PatchGAN [13]. We replaced the Eqs. (1) and (2) by least-squares loss as in [14]. These settings follows [5,6]. The CycleGAN was trained using Adam [15] for the first 1×10^5 iterations at fixed learning rate of 0.0002, and the last 1×10^5 iterations at learning rate which linearly reducing to zero. The balancing weights were empirically determined as $\lambda_{Cycle} = 3$ and $\lambda_{GC} = 0.3$. CT and MR volumes are normalized such that intensity of [−150, 350] HU and [0, 100] are mapped to [0, 255], respectively.

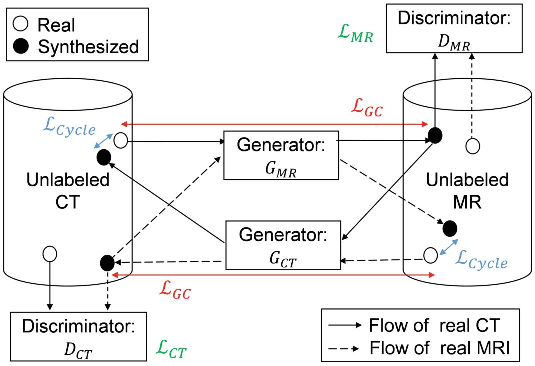


Fig. 2. Workflow of the proposed method. G_{CT} and G_{MR} are generator networks that translate MR to CT images, and CT to MR images, respectively. D_{CT} and D_{MR} are discriminator networks to distinguish between real and synthesized images. The cycle consistency loss $\mathcal{L}_{Cycle}$ is a regularization term defined by the difference between real and reconstructed image. To improve the accuracy at the edges, loss function is regularized by gradient consistency loss $\mathcal{L}_{GC}$.

3 Result

3.1 Quantitative Evaluation on Image Synthesis

To evaluate image synthesis, we investigated dependency of the accuracy on the number of training data and with or without the GC loss. The CycleGAN was trained with datasets of different sizes, (i) 20 MR and 20 CT volumes, (ii) 302 MR and 613 CT volumes, and both with and without GC loss. We conducted two experiments. The first experiment used three sets of paired MR and CT volumes of the same patient for test data. Because availability of paired MR and CT volumes was limited, we conducted the second experiment in which unpaired 10 MR and 20 CT volumes were used.

In the first experiment, we evaluated synthesized CT by means of mean absolute error (MAE) and peak-signal-to-noise ratio (PSNR) [dB] between synthesized CT and ground truth CT, both of which were normalized as mentioned in 2.2. The ground truth CT here is a CT registered to the MR of the same patient. CT and MR volumes were aligned using landmark-based registration as initialization, and then aligned using rigid and non-rigid registration. The results of MAE and PSNR are shown in Table 1. PSNR is calculated as $PSNR = 20 \log_{10} \frac{255}{\sqrt{MSE}}$, where MSE is mean squared error. The average of MAE decreased and PSNR increased according to the increase of training data size and inclusion of GC loss, respectively. Figure 3 shows representative results.

In the second experiment, we tested with unpaired 10 MR and 20 CT volumes. Mutual information (MI) between synthesized CT and original MR was used for evaluation when the paired ground truth was not available. The quantitative results are show in Fig. 4(a). The left side is the box and whisker plots

Table 1. Mean absolute error (MAE) and peak-signal-to-noise ratio (PSNR) between synthesized and real CT volumes.

		20 volumes		>300 volumes	
		w/o GC	/w GC	w/o GC	/w GC
MAE	Patient #1	30.121	30.276	26.899	26.388
	Patient #2	26.927	26.911	22.319	21.593
	Patient #3	33.651	32.155	29.630	28.643
	Average ± SD	30.233 ± 2.177	29.781 ± 1.777	26.283 ± 1.367	25.541 ± 1.129
PSNR	Patient #1	14.797	14.742	15.643	15.848
	Patient #2	15.734	15.628	17.255	17.598
	Patient #3	14.510	14.820	15.674	15.950
	Average ± SD	15.014 ± 0.330	15.063 ± 0.380	16.190 ± 0.273	16.465 ± 0.296

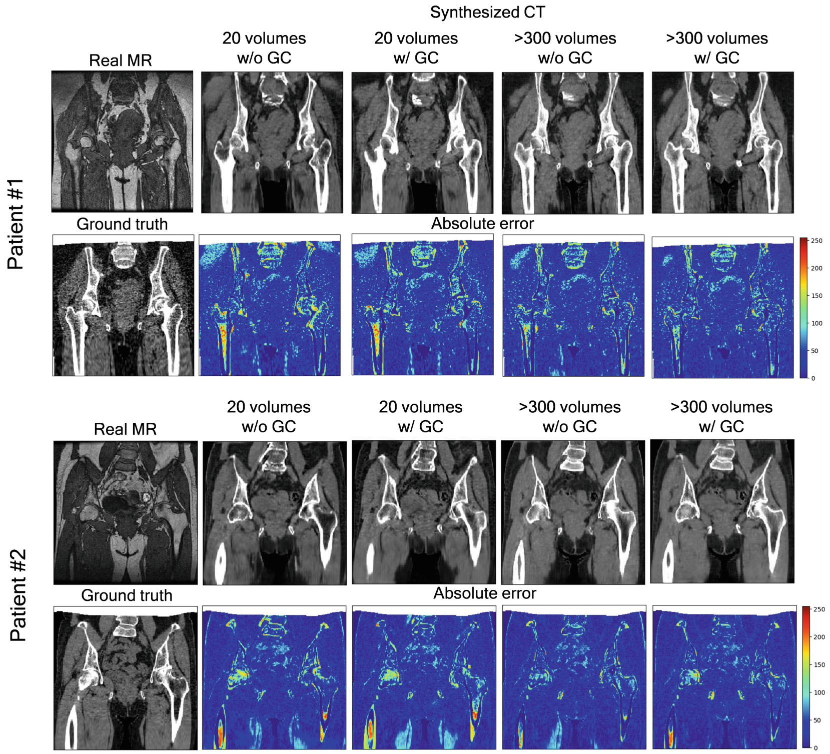
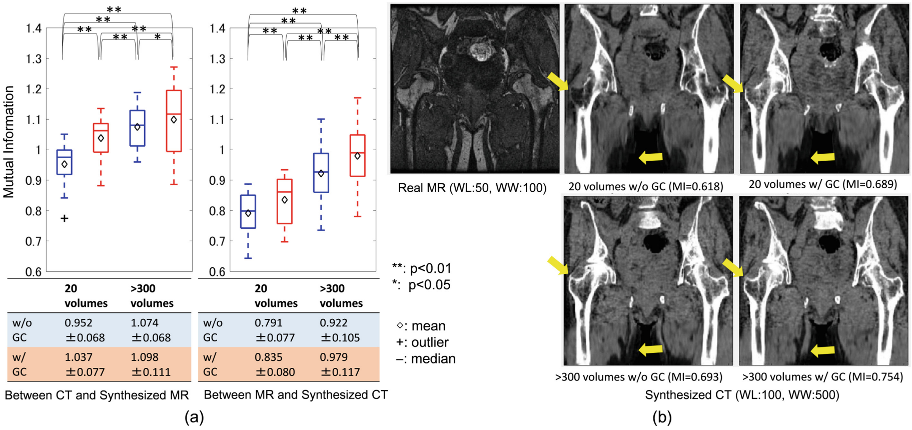


Fig. 3. Representative results of the absolute error between the ground truth paired CT and synthesized CT from two patients. Since the FOV of MR and CT volumes are slightly different, there is no corresponding region near the top edge of the ground truth volumes (filled with white color). This area was not used for evaluation.



	20 volumes	>300 volumes
w/o GC	0.952 ±0.068	1.074 ±0.068
w/ GC	1.037 ±0.077	1.098 ±0.111

Between CT and Synthesized MR

	20 volumes	>300 volumes
w/o GC	0.791 ±0.077	0.922 ±0.105
w/ GC	0.835 ±0.080	0.979 ±0.117

Between MR and Synthesized CT

Fig. 4. Evaluation of similarity between the real and synthesized volumes. (a) quantitative comparison of mutual information on different training data size with and without the gradient-consistency loss. (b) representative result of one patient.

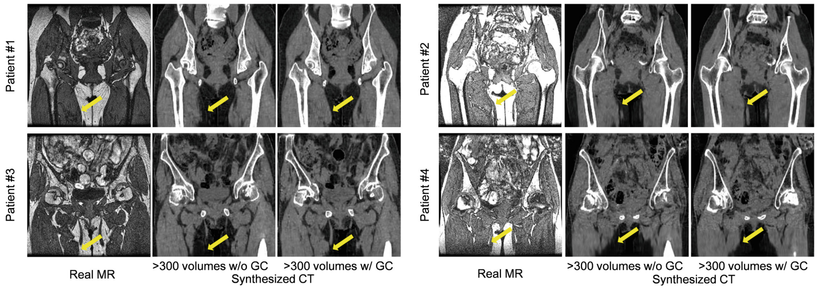
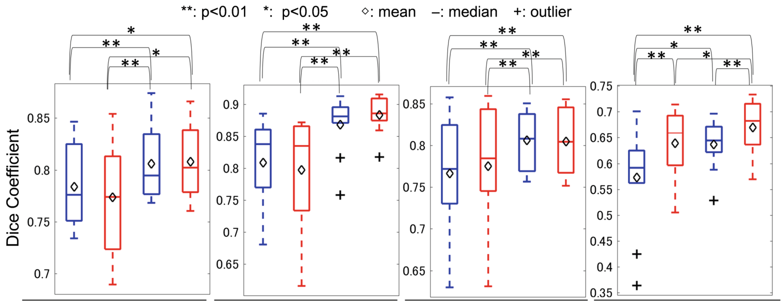


Fig. 5. Representative results of translation from real MR to synthesized CT of four patients with and without the gradient consistency loss. As indicated by arrows, synthesized volumes with gradient consistency loss helped to preserve the shape near the adductor muscles.

of the mean of each slice of MI between real CT and synthesized MR (i.e., 20 data points in total). The right side is the mean of MI between real MR and synthesized CT (i.e., 10 data points in total). The result shows that the larger number of training data yielded statistically significant improvement ($p < 0.01$) according to the paired t-test in MI. The GC loss also leads to an increase in MI between MR and synthesized CT ($p < 0.01$). Figures 4(b) and 5 show examples of the visualization of real MR and synthesized CT volumes. As indicated by arrows, we can see that synthesized volumes with GC loss preserved the shape near the femoral head and adductor muscles.



	20 volumes	>300 volumes
w/o GC	0.784 ±0.041	0.806 ±0.036
w/ GC	0.774 ±0.055	0.808 ±0.036

(a) Pelvis

	20 volumes	>300 volumes
w/o GC	0.809 ±0.073	0.868 ±0.047
w/ GC	0.797 ±0.085	0.883 ±0.029

(b) Femur

	20 volumes	>300 volumes
w/o GC	0.767 ±0.066	0.806 ±0.036
w/ GC	0.775 ±0.070	0.804 ±0.040

(c) Gluteus medius

	20 volumes	>300 volumes
w/o GC	0.573 ±0.106	0.637 ±0.049
w/ GC	0.639 ±0.067	0.669 ±0.054

(d) Gluteus minimus

Fig. 6. Evaluation of segmentation accuracy on different training data size in CycleGAN with and without the gradient-consistency loss. Segmentation of (a) pelvis, (b) femur, (c) gluteus medius and (d) gluteus minimus muscle in MR volumes were performed using MR-to-CT synthesis.

3.2 Quantitative Evaluation on Segmentation

To demonstrate the applicability of image synthesis in segmentation task, we evaluated the segmentation accuracy. Twenty labeled CT datasets were used to train the segmentation network. Then, we evaluated the segmentation accuracy with 10 MR volumes with manual segmentation labels of the gluteus medius and minimus muscles and femur.

We employed the 2D U-net proposed by Ronneberger et al. [16] as segmentation network, which is widely used in medical image analysis and demonstrated high performance with a limited number of labeled volumes. In MRI, muscle boundaries are clearer while bone boundaries are clearer in CT. To incorporate the advantage of both CT and MR, we modified the 2D U-net to take the two-channel input of both CT and synthesized MR images. We trained on 2D U-net using Adam [15] for 1×10^5 iterations at learning rate of 0.0001. At the test phase, a pair of MR and synthesized CT was used as two-channel input.

The results with 4 musculoskeletal structures for 10 patients are shown in Fig. 6 (i.e., 10 data points in total on each plot). The result shows that the larger number of training data yielded statistically significant improvement in DICE on pelvis ($p < 0.01$), femur ($p < 0.01$), glutes medius ($p < 0.01$) and glutes minimus regions ($p < 0.05$) of paired t-test. The GC loss also leads to an increase in DICE on the glutes minimus regions ($p < 0.01$). The average DICE coefficient in the cases trained with more than 300 cases and GC loss was 0.808 ± 0.036 (pelvis), 0.883 ± 0.029 (femur), 0.804 ± 0.040 (gluteus medius) and 0.669 ± 0.054 (gluteus minimus), respectively. Figure 7 shows example visualization of real MR, synthesized CT, and estimated label for one patient. The result with GC loss has smoother segmentation not only in the gluteus minimus but also near the adductor muscles.

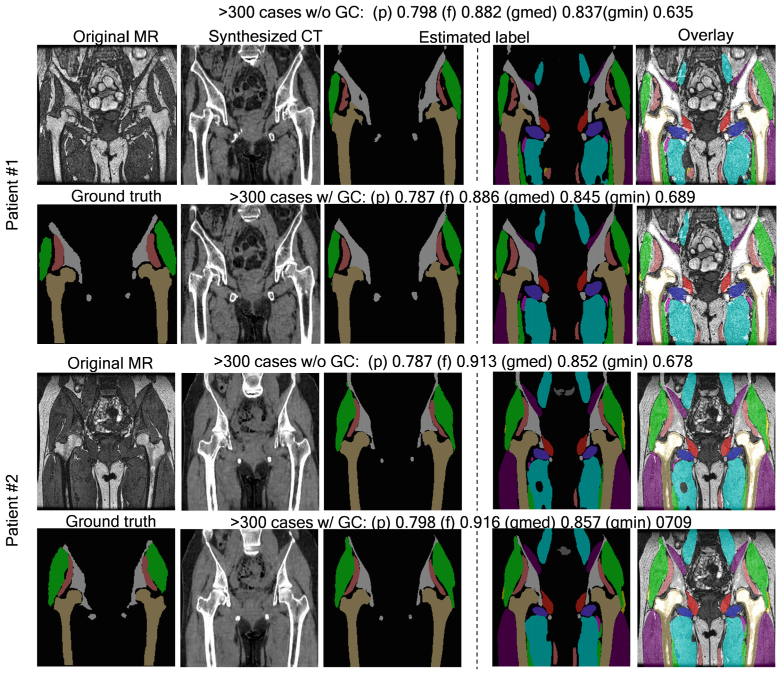

>300 cases w/o GC: (p) 0.798 (f) 0.882 (gmed) 0.837(gmin) 0.635
Original MR
Synthesized CT
Estimated label
Overlay
Patient #1
Ground truth
>300 cases w/ GC: (p) 0.787 (f) 0.886 (gmed) 0.845 (gmin) 0.689
Original MR
>300 cases w/o GC: (p) 0.787 (f) 0.913 (gmed) 0.852 (gmin) 0.678
Patient #2
Ground truth
>300 cases w/ GC: (p) 0.798 (f) 0.916 (gmed) 0.857 (gmin) 0709

Fig. 7. Representative results of segmentation from one patient. The ground truth label is consist of 4 musculoskeletal structures in MRI. Although we evaluated only on 4 structures because ground truth were not available for the other structures on MRI, all 22 estimated labels are shown for qualitative evaluation. In the right-most column, all estimated labels are overlayed on the real MRI. p, f, gmed, gmin denote DICE of pelvis, femur, gluteus medius, and gluteus minimus, respectively.

4 Discussion and Conclusion

In this study, we proposed an image synthesis method which extended the CycleGAN approach by adding the GC loss to improve the accuracy at the boundaries. Specifically, the contributions of this paper are (1) introduction of GC loss in CycleGAN, and (2) quantitative and qualitative evaluation of the dependency of both image synthesis accuracy and segmentation accuracy on a large number of training data. One limitation in this study is that we excluded the patients with implants, while our target cohort (i.e., THA patients) sometime has implant on one side, for example, in case of the planning of secondary surgery. As a comparison against a single modality training, we performed 5-fold cross validation of MR segmentation using 10 labeled MR volumes (i.e., trained with 8 MR volumes and tested on remaining 2 MR volumes) using U-net segmentation network. The DICE was 0.815 ± 0.046 (pelvis), 0.921 ± 0.023 (femur), 0.825 ± 0.029 (gluteus medius) and 0.752 ± 0.045 (gluteus minimus), respectively. We found the gap of accuracy between modality independent and dependent segmentation. A potential improvement of modality independent segmentation is to construct an end-to-end network that performs image synthesis and segmentation [17]. Our future work also includes development of a method that effectively incorporates information in unlabeled CT and MR volumes to improve segmentation accuracy [18].

References

1. Cvitanic, O.: MRI diagnosis of tears of the hip abductor tendons (gluteus medius and gluteus minimus). Am. J. Roentgenol. **182**(1), 137–143 (2004)
2. Torrado-Carvajal, A.: Fast patch-based pseudo-CT synthesis from T1-weighted MR images for PET/MR attenuation correction in brain studies. J. Nuclear Med. **57**(1), 136–143 (2016)
3. Zhao, C., Carass, A., Lee, J., He, Y., Prince, J.L.: Whole brain segmentation and labeling from CT using synthetic MR images. In: Wang, Q., Shi, Y., Suk, H.-I., Suzuki, K. (eds.) MLMI 2017. LNCS, vol. 10541, pp. 291–298. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67389-9_34
4. Kamnitsas, K., et al.: Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. In: Niethammer, M. (ed.) IPMI 2017. LNCS, vol. 10265, pp. 597–609. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-59050-9_47
5. Zhu, J.Y., et al.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2223–2232 (2017)
6. 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
7. Gilles, B.: Musculoskeletal MRI segmentation using multi-resolution simplex meshes with medial representations. Med. Image Anal. **14**(3), 291–302 (2010)
8. Ranzini, M.B.M., et al.: Joint multimodal segmentation of clinical CT and MR from hip arthroplasty patients. In: Glocker, B., Yao, J., Vrtovec, T., Frangi, A., Zheng, G. (eds.) MSKI 2017. LNCS, vol. 10734, pp. 72–84. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-74113-0_7
9. Hamarneh, G., Jassi, P., Tang, L.: Simulation of ground-truth validation data via physically-and statistically-based warps. In: Metaxas, D., Axel, L., Fichtinger, G., Székely, G. (eds.) MICCAI 2008. LNCS, vol. 5241, pp. 459–467. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-85988-8_55
10. Tustison, N.J.: N4ITK: improved N3 bias correction. IEEE Trans. Med. Imaging **29**(6), 1310–1320 (2010)
11. Penney, G.P.: A comparison of similarity measures for use in 2-D-3-D medical image registration. IEEE Trans. Med. Imaging **17**(4), 586–595 (1998)
12. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9906, pp. 694–711. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_43
13. Isola, P., et al.: Image-to-image translation with conditional adversarial networks. arXiv preprint (2017)
14. Mao, X., et al.: Multi-class generative adversarial networks with the L2 loss function. CoRR, abs/1611.04076 2 (2016)

15. Kingma, D.P., et al.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
16. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
17. Huo, Y., et al.: Adversarial synthesis learning enables segmentation without target modality ground truth. arXiv preprint arXiv:1712.07695 (2017)
18. Zhang, Y., Yang, L., Chen, J., Fredericksen, M., Hughes, D.P., Chen, D.Z.: Deep adversarial networks for biomedical image segmentation utilizing unannotated images. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10435, pp. 408–416. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66179-7_47