Published in final edited form as:

Deep Learn Med Image Anal Multimodal Learn Clin Decis Support (2018). 2018 September ; 11045: 155–164. doi:10.1007/978-3-030-00889-5_18.

Unpaired Deep Cross-Modality Synthesis with Fast Training

Lei Xiang^{#1}, Yang Li^{#2,3}, Weili Lin³, Qian Wang¹, and Dinggang Shen³

¹Institute for Medical Imaging Technology, School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China

²Department of Computer Science, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

³Department of Radiology and BRIC, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

[#] These authors contributed equally to this work.

Abstract

Cross-modality synthesis can convert the input image of one modality to the output of another modality. It is thus very valuable for both scientific research and clinical applications. Most existing cross-modality synthesis methods require large dataset of paired data for training, while it is often non-trivial to acquire perfectly aligned images of different modalities for the same subject. Even tiny misalignment (i.e., due patient/organ motion) between the cross-modality paired images may place adverse impact to training and corrupt the synthesized images. In this paper, we present a novel method for cross-modality image synthesis by training with the unpaired data. Specifically, we adopt the generative adversarial networks and conduct the fast training in cyclic way. A new structural dissimilarity loss, which captures the detailed anatomies, is introduced to enhance the quality of the synthesized images. We validate our proposed algorithm on three popular image synthesis tasks, including brain MR-to-CT, prostate MR-to-CT, and brain 3T-to-7T. The experimental results demonstrate that our proposed method can achieve good synthesis performance by using the unpaired data only.

1 Introduction

Due to the complementary information contained in different imaging modalities (e.g., CT images, T1- and T2-weighted MR images), multi-modal images are usually captured and fused for disease diagnosis, treatment planning, etc. However, acquisition of multimodal images can be time-consuming and costly. Furthermore, the fusion often requires accurate cross-modality registration and can be degraded by the deformation of the organs.

Cross-modality synthesis is thus valuable for both scientific research and clinical application. Although each modality presents different characteristic of the underlying anatomy, individual modalities are highly correlated when scanning the same anatomical

structure and revealing the tissue appearance from different perspectives. Thus, synthesizing images of one modality based on the images of another modality is theoretically possible. However, the mapping between the two different modalities are highly nonlinear, which makes the synthesis task difficult to accomplish.

Over the past few years, various methods have been proposed for cross-modality medical image synthesis. Typical works include coupled sparse representation [1] and deep convolutional neural networks [2–4]. These methods usually require paired data for training, i.e., well-aligned source and target modalities from the same subject. However, it is not always easy to get the perfectly paired data, which thus strongly limits the application of cross-modality synthesis. Moreover, misalignment within the paired source/target data is sometimes inevitable (though tiny), and it could cause ambiguity or even devastate current synthesis methods.

Unsupervised synthesis has already been explored in [5], which only requires unpaired data for training. They used cross-modality nearest neighbor search to produce the candidate for each target voxel, then simultaneously maximized the global mutual information between candidate and source images. Local spatial consistency was enforced to generate the final target image. The performance of the method is highly dependent on the accuracy of the nearest neighbor searching.

Recently, unsupervised deep learning models have been applied for image synthesis. Cycle-GAN [6], for example, has been used to synthesize CT from MR [7]. However, it is insufficient to simply borrow the Cycle-GAN model while many properties of the medical images are ignored. We argue that the synthesis of medical images is quite different from natural images due to the 3D nature of many medical image modalities. Thus, in this work, we train the deep network in a quasi-3D way and design a 3D structural dissimilarity loss for several popular medical tasks. Particularly, inspired by the structural similarity metric (SSIM), we introduce a new structural dissimilarity loss to improve the boundary contrast of the synthesized image.

We also simplify the generator in GAN to decrease the number of the parameters, which leads to faster training yet better synthesis quality. Our generator combines the advantages of Unet [8] and deep residual net [9], and is termed as Res-Unet. Our simplified model can then be well trained within 3 h. We conduct abundant experiments to verify the promising performances of our method. Specifically, we perform brain MR-to-CT synthesis, prostate MR-to-CT synthesis and brain 3T-to-7T MR synthesis, respectively. Several examples of our datasets are shown in Fig. 1, where the differences between the paired and the unpaired data are clear. Note that in this paper we use the unpaired data only for all the experiments.

2 Method

2.1 Loss Design

We aim to accomplish the cross-modality synthesis by the Cycle-Consistent Adversarial Networks. Suppose we have two modality images X and Y . Then, the goal of our method is

to learn the mapping function between these two modalities. We define the training samples as $\{x_i\}_{i=1}^N \in X$ and $\{y_j\}_{j=1}^M \in Y$. As illustrated in Fig. 2(a), there are two mapping functions, i.e., $G: X \rightarrow Y$ and $F: Y \rightarrow X$ in this cross-modality synthesis task. The two mapping functions can be modeled by deep neural networks. Besides, two adversarial discriminators D_X and D_Y are trained, such that D_X tries to distinguish real images $\{x_i\}$ and the synthesized images $\{F(y_j)\}$. Similarly, D_Y tries to distinguish $\{y_j\}$ and $\{G(x_i)\}$. In order to quantify the variation of the anatomical structures between the real images and the synthesized images, we also introduce the new structural dissimilarity loss. Therefore, the objective of the network as shown in Fig. 2 (a) mainly contains three terms: the adversarial loss ($\mathcal{L}_{GAN}$), the cycle consistency loss ($\mathcal{L}_{CYC}$) and the structural dissimilarity loss ($\mathcal{L}_{DSSIM}$):

$$\begin{aligned} \mathcal{L}(G, F, D_X, D_Y) = & \mathcal{L}_{GAN}(G, D_Y, X, Y) + \mathcal{L}_{GAN}(F, D_X, Y, X) \\ & + \lambda \mathcal{L}_{CYC}(G, F) + \beta \mathcal{L}_{DSSIM}(G, F), \end{aligned} \quad (1)$$

where λ and β control the relative importance of individual loss terms. We set $\lambda = 10$ and set $\beta = 1$ in this work.

Adversarial Loss.—Adversarial loss is applied to both mapping functions G and F . For the mapping function $G: X \rightarrow Y$ and its corresponding discriminator D_Y , the objective function is expressed as:

$$\mathcal{L}_{GAN}(G, D_Y, X, Y) = \mathbb{E}_{y \sim P_{data}(y)} [\log D_Y(y)] + \mathbb{E}_{x \sim P_{data}(x)} [\log(1 - D_Y(x))] \quad (2)$$

G intends to generate the target modality image $G(x)$ that appears to be similar to real target image (Y), while D_Y aims to distinguish whether the input to the discriminator is the synthesized image $G(x)$ or a real image $y \in Y$. Therefore, G tries to minimize this objective function while the adversarial D tries to maximize it, i.e.

$G^* = \operatorname{argmin}_G \max_{D_Y} \mathcal{L}_{GAN}(G, D_Y, X, Y)$. Similar adversarial loss is also applied for the mapping function $F: Y \rightarrow X$: i.e., $F^* = \operatorname{argmin}_F \max_{D_X} \mathcal{L}_{GAN}(F, D_X, Y, X)$

Cycle Consistency Loss.—To further reduce the ambiguity in solving the mapping functions, we enforce the cycle-consistency constraint, which means the difference between the input modality image and the cyclically synthesized image should be minimized. The illustration for the cycle consistency loss is shown in Fig. 2(b) and (c) for both synthesis direction, i.e., $x \rightarrow G(x) \rightarrow F(G(x))$ should be similar with x and $y \rightarrow G(y) \rightarrow F(G(y))$ should be similar with y . This cycle-consistency loss can thus be defined as:

$$\mathcal{L}_{cyc}(G, F) = E_{x \sim P_{data}(x)}[\|F(G(x)) - x\|_1] + E_{y \sim P_{data}(y)}[\|F(G(y)) - y\|_1]. \quad (3)$$

Structural Dissimilarity Loss.—As the global L1 loss focuses on the entire image space, it ignores many local structural details. Structural information is usually critical in medical images as they are closely related to delineating the boundaries of tissues and organs. In order to further improve the quality of the synthesized images regarding anatomical details, we propose to take advantage of SSIM to restore the local structures in the synthesized image. This leads to the new structural dissimilarity loss (DSSIM), which is a distance metric extended from SSIM:

$$\begin{aligned} \mathcal{L}_{DSSIM}(G, F) = E_{x \sim P_{data}(x)} & \left[\frac{1 - SSIM(x, F(G(x)))}{2} \right] \\ & + E_{y \sim P_{data}(y)} \left[\frac{1 - SSIM(y, F(G(y)))}{2} \right]. \end{aligned} \quad (4)$$

2.2 Architecture of the Generator/Discriminator

There are two networks in the Cycle-Consistent Adversarial Networks, i.e., the generator and the discriminator. The generator, which is critical to the quality of the generated images, has many layers with abundant parameters, making the training process very slow. In order to design a more efficient network, we intend to take advantages from two popular architectures, i.e., Unet [8] and deep residual network [9].

Unet is widely used in many medical image analysis researches, as it has shown promising results on various tasks, including image segmentation and image synthesis. Unet consists of an encoding path and a decoding path, with skip connection in each corresponding level. This design ensures the network to have a large receptive field to capture both local and global image appearances. Salient high-level features can thus be extracted, which is essential to the cross-modality mapping trained with unpaired samples. Deep residual network, is also adopted in many research tasks, such as image classification and image super-resolution. The most important component of deep residual network is the residual block, which consists of two convolutional layers with an identity mapping, as shown in Fig. 3. The residual block is designed to alleviate the gradient vanishing issue; in the meantime, it can also boost information exchange across different layers. Inspired by the two networks, we design a deep network called Res-Unet in this work. Our network fuses advantages of Unet and the residual block, as its architecture is illustrated in Fig. 3. There are 2 pooling stages, 2 deconvolution stages and 5 residual blocks in our generator.

3 Experimental Results

3.1 Datasets and Tasks

We utilize three real datasets to evaluate our cross-modality synthesis method. The datasets and the relative synthesis tasks are introduced below.

1. **Brain MR-to-CT dataset.** This dataset consists of 16 subjects, each of whom comes with an MR and a CT scan. The voxel sizes of the CT and MR images are $0.59 \times 0.59 \times 0.59 \text{ mm}^3$ and $1.2 \times 1.2 \times 1 \text{ mm}^3$, respectively. We separate the 16 subjects into a training set containing 10 subjects and a testing set containing 6 subjects.
2. **Prostate MR-to-CT dataset.** The prostate dataset consists of 22 subjects. The voxel sizes of the CT and MR images are $1.17 \times 1.17 \times 1 \text{ mm}^3$ and $1 \times 1 \times 1 \text{ mm}^3$, respectively. We also separate the 22 subjects into two parts: a training set containing 14 subjects and a testing set containing 8 subjects.
3. **Brain 3T-to-7T dataset.** This dataset consists of 15 subjects. The voxel sizes of the 3T MR and 7T MR are $1 \times 1 \times 1 \text{ mm}^3$ and $0.65 \times 0.65 \times 0.65 \text{ mm}^3$, respectively. These 15 subjects are separated into a training set containing 10 subjects and a testing set containing 5 subjects.

For both brain and prostate MR-to-CT tasks, the CT images are linearly aligned (by FLIRT in FSL) to the corresponding MR images and resampled to the same size of the MR images. For brain 3T-to-7T dataset, corresponding 3T and 7T images are also linearly aligned. The nonlinear deformations between images in the same subject are left there. The intensities are normalized to $[0, 1]$ in each image.

3.2 Implementation Details

In this paper, PyTorch implementation for the basic Cycle-GAN [6] is used in all the experiments¹. The generator is replaced by the proposed Res-Unet in our method. In the training phase, we extract consecutive 2D axial slices from the 3D image as the training samples. The training samples from two different input modalities are drawn separately, such that the samples in each pair are totally independent. The sampling process results in the *unpaired* training dataset, which allows no alignment of the training images in practical usage. Horizontal flipping is used to augment the training datasets. We apply Adam optimization with momentum of 0.9 and perform 100 epochs in the training stage. The batch size is set to 1 and the initial learning rate is set to 0.0002. To quantitatively evaluate the results, we use the commonly accepted metrics of peak signal-noise ratio (PSNR), normalized mean squared error (NMSE) and structural similarity (SSIM). In general, higher PSNR, lower NMSE and high SSIM indicate better perceptive quality of the synthesis result.

¹<https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix>.

3.3 Quantitative and Visual Comparisons

In this section, we compare the cross-modality synthesis results by our proposed method and the previously reported Cycle-GAN model. Comparisons are conducted on all three tasks: brain MR-to-CT, prostate MR-to-CT, and brain 3T-to-7T. First, we show the effectiveness of the proposed new generator Res-UNet. Comparisons between ‘Basic Cycle-GAN’ and ‘Res-UNet’ have been summarized in Table 1. We can see that with the new generator the synthesis results are improved on all three datasets, which demonstrates the superiority of the proposed generator. Moreover, with the new DSSIM loss added, the synthesis performance is further improved. In general, the quantitative results in Table 1 show that our proposed method (‘Res-UNet + DSSIM’) achieves best results on all three tasks, in terms of all evaluation metrics of PSNR, NMSE and SSIM.

To give an intuitive view, visualization of the synthesized results using ‘Basic Cycle-GAN’ and the proposed ‘Res-UNet + DSSIM’ are presented in Figs. 4, 5 and 6 for prostate MR-to-CT, brain MR-to-CT, and brain 3T-to-7T, respectively. Compared to ‘Basic Cycle-GAN’, ‘Res-UNet + DSSIM’ can obtain better synthesis results with clearer tissue/organ boundaries. For example, in the coronal view of the prostate MR-to-CT task in Fig. 4, the two bones in the hip joint are successfully separated and synthesized by ‘Res-UNet + DSSIM’, while the boundaries of the bones appear blur in the synthesized result by ‘Basic Cycle-GAN’ (orange box in the figure). Also we could observe that the anatomical details pointed by the red and green boxes are clearer in ‘Res-UNet + DSSIM’. Similar observations can be found on brain MR-to-CT and brain 3T-to-7T dataset in Figs. 5 and 6.

Meanwhile, note that our training is based on 3 consecutive axial slices, while the synthesized results are consistent on all three views. That is, our method can well handle the synthesis task of 3D medical images, even though only 3 consecutive slices are used in training. In the testing stage, we process every 3 axial slices each time, while the final output of the 3D volume can be obtained by averaging all synthesis results.

We conduct another experiment on brain MR-to-CT dataset to show the fast convergence of our ‘Res-UNet + DSSIM’ compared to ‘Basic Cycle-GAN’. We show the synthesis result by training with same epoch. The results are shown in Fig. 7. We can see that with same training epoch, our proposed model gets better result. It takes 3 h to train our model, while Basic Cycle-GAN takes 17 h. And our model contains 13.3 M parameters, which is 1/4 of ‘Basic Cycle-GAN’. That means our proposed model could train fast with less parameters, but achieve best synthesis result. The testing time is 6 s for a 3D image of size $181 \times 234 \times 149$.

4 Conclusion

We proposed a novel Res-UNet architecture as the generator and solve cross-modality image synthesis by GAN. In particular, we accomplish the synthesis tasks on three different scenarios by training with the unpaired data, which indicates that our method has great potentials to many real clinical applications. This Res-UNet generator, which benefits from the novel loss design, has shown its superior performances by mapping between different

images modalities with large appearance variation. In our future work, we will conduct large-scale evaluation in clinical applications, and demonstrate that the proposed image synthesis technique can be used as a new tool to reshape multi-modal image fusion and subsequent analysis.

References

1. Cao T, Zach C, Modla S, Powell D, Czymmek K, Niethammer M: Multi-modal registration for correlative microscopy using image analogies. *Med. Image Anal* 18, 914–926 (2014) [PubMed: 24387943]
2. Xiang L, et al.: Deep embedding convolutional neural network for synthesizing CT image from T1-Weighted MR image. *Med. Image Anal* 47, 31–44 (2018) [PubMed: 29674235]
3. Nie D, et al.: Medical image synthesis with context-aware generative adversarial networks. In: Descoteaux M, Maier-Hein L, Franz A, Jannin P, Collins DL, Duchesne S (eds.) MICCAI 2017. LNCS, vol. 10435, pp. 417–425. Springer, Cham (2017). 10.1007/978-3-319-66179-7_48
4. Xiang L, et al.: Deep auto-context convolutional neural networks for standard-dose PET image estimation from low-dose PET/MRI. *Neurocomputing* 267, 406–416 (2017) [PubMed: 29217875]
5. Vemulapalli R, Van Nguyen H, Kevin Zhou S: Unsupervised cross-modal synthesis of subject-specific scans. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 630–638 (2015)
6. Zhu J-Y, Park T, Isola P, Efros AA: Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint arXiv:1703.10593* (2017)
7. Wolterink JM, Dinkla AM, Savenije MHF, Seevinck PR, van den Berg CAT, Išgum I: Deep MR to CT synthesis using unpaired data. In: Tsaftaris SA, Gooya A, Frangi AF, Prince JL (eds.) SASHIMI 2017. LNCS, vol. 10557, pp. 14–23. Springer, Cham (2017). 10.1007/978-3-319-68127-6_2
8. Ronneberger O, Fischer P, Brox T: U-Net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). 10.1007/978-3-319-24574-4_28
9. He K, Zhang X, Ren S, Sun J: Deep Residual Learning for Image Recognition. *arXiv preprint arXiv: 1512.03385* (2015)

Fig. 1.

Examples of the paired (top) and unpaired training data (bottom) for three tasks: brain MR-to-CT, prostate MR-to-CT, and brain 3T-to-7T MR. In the paired data, the input images (X and Y) belong to the same subject and registered. In the unpaired data, the inputs images are clearly misaligned.

Fig. 2.

The Cycle-Consistent Adversarial Networks (a) used for cross-modality synthesis are illustrated in (a). There are two cycle mappings as in (b) and (c).

Fig. 3. Illustration of the proposed Res-Unet architecture as the generator.

Fig. 4.

Visual comparison for the prostate MR-to-CT synthesis task. (Color figure online)

Fig. 5.
Visual comparison for the brain MR-to-CT synthesis task. (Color figure online)

Fig. 6.

Visual comparison for the brain 3T-to-7T synthesis task. (Color figure online)

Fig. 7.
Visual comparison for Basic Cycle-GAN and proposed method during training.

Table 1.

Comparisons of the synthesis results by different methods.

	Brain MR-to-CT			Prostate MR-to-CT			Brain 3T-to-7T		
	PSNR	NMSE	SSIM	PSNR	NMSE	SSIM	PSNR	NMSE	SSIM
Basic cycle-GAN	24.89	0.0569	0.896	30.27	0.0306	0.931	26.71	0.0461	0.918
Res-Unet	25.41	0.0532	0.904	30.58	0.0303	0.929	27.21	0.0436	0.922
Res-Unet + DSSIM (Proposed method)	<u>25.62</u>	<u>0.0525</u>	<u>0.909</u>	<u>31.23</u>	<u>0.0277</u>	<u>0.937</u>	<u>27.52</u>	<u>0.0421</u>	<u>0.926</u>