Disentangled Representation Learning For Deep MR To CT Synthesis Using Unpaired Data

Conference Paper · September 2021

DOI: 10.1109/ICIP42928.2021.9506660

2 authors, including:

Runze Wang
Shanghai Jiao Tong University
10 PUBLICATIONS **37** CITATIONS

SEE PROFILE

DISENTANGLED REPRESENTATION LEARNING FOR DEEP MR TO CT SYNTHESIS USING UNPAIRED DATA

Runze Wang *Guoyan Zheng**

Institute of Medical Robotics, School of Biomedical Engineering, Shanghai Jiao Tong University, China
*Correspondence: guoyan.zheng@sjtu.edu.cn

2021 IEEE International Conference on Image Processing (ICIP) | 978-1-6654-4115-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICIP42928.2021.9506660

ABSTRACT

Many different methods have been proposed for generation of synthetic CT (sCT) from MR images. Most of these methods depend on paired-wise aligned MR and CT training images of the same patient, which are difficult to obtain. In this paper, we propose a novel disentangled representation learning method for MR to CT synthesis using unpaired data. Specifically, we first embed images onto two spaces: a modality-invariant geometry space capturing the shared anatomical information across different imaging domains, and a modality-specific appearance space. From the embedding, a sCT image can be synthesized from a MR image by taking the encoded geometry features from the MR image and an appearance vector sampled from the appearance space of a CT image. To handle the challenging of distinguishing cortical bone from air in MR images, where both of them have low intensity values, we propose a novel Geometry Similarity Module (GSM) to take the context information into consideration. Experimental results demonstrated that our approach achieved better or equivalent results than the state-of-the-art.

Index Terms— MR-to-CT synthesis, Disentangled representation learning, Geometry similarity, Context information

1. INTRODUCTION

Despite the fact that Computed Tomography (CT) images have limited soft tissue contrast and result in extra radiation to the patients, CT imaging is critical for various applications, e.g., radiotherapy treatment planning and Positron Emission Tomography (PET) attenuation correction. This is because CT images offer accurate presentation of patient geometry and more importantly, CT values in Hounsfield units (HU), which measure tissue attenuation coefficients, can be directly converted to electron density for radiation dose calculation. Recently, interests in replacing CT with magnetic resonance imaging (MRI) have grown rapidly due to MRI's free of ionizing radiation, excellent soft tissue contrast, and ability of multiparametric imaging through various MRI sequences. The main challenges in replacing CT with MRI, however, are a) the MRI intensity values, unlike CT values, are not directly related to electron densities; and b) conventional MRI sequences pose dramatic limitations for distinguishing cortical bone from air. It is therefore desirable to have a method to derive CT-equivalent information from MR images. Such MR-based CT-equivalent data are often referred to as synthetic CT (sCT) in the literature.

Many different methods have been proposed for generation of synthetic CT from MR images [1, 2, 3, 4, 5, 6, 7, 8]. Most of these methods depend on pairwise aligned MR and CT training images of the same patient, which are difficult to obtain. Any error in aligning MR and CT images could lead to errors in generating sCT. Inspired by the work of [9], several groups have developed methods for automated MR-to-CT synthesis using cycle-consistent Generative Adversarial Networks (CycleGAN), which could be trained without the need for paired training data [5, 8, 10, 11]. One major limitation of CycleGAN is that it only learns one-to-one mappings, i.e., the model associated each input with a single output image. We believe that the relationships between MR domain and CT domain are more complex, and better characterized as many-to-many, given the fact that there are many factors influencing the image generation process of these two different imaging modalities.

In this paper, we propose a novel approach for MR to CT synthesis using unpaired data. Specifically, we first embed images onto two spaces: a modality-invariant geometry space capturing the shared anatomical information across two different imaging domains, and a modality-specific appearance space. From the embedding, a sCT image can be synthesized from a MR image by taking the encoded geometry features from the MR image and an appearance vector sampled from the appearance space of a CT image. Our contributions can be summarized as follows:

1. We introduce a disentangled representation learning method for MR to CT synthesis using unpaired data.
2. To facilitate the disentangled learning when 2D slices are used, we introduce a slice corresponding strategy.
3. Furthermore, to handle the challenging of distinguishing cortical bone from air in MR images, where both

The work was partially supported by the National Natural Science Foundation of China via project U20A20199.

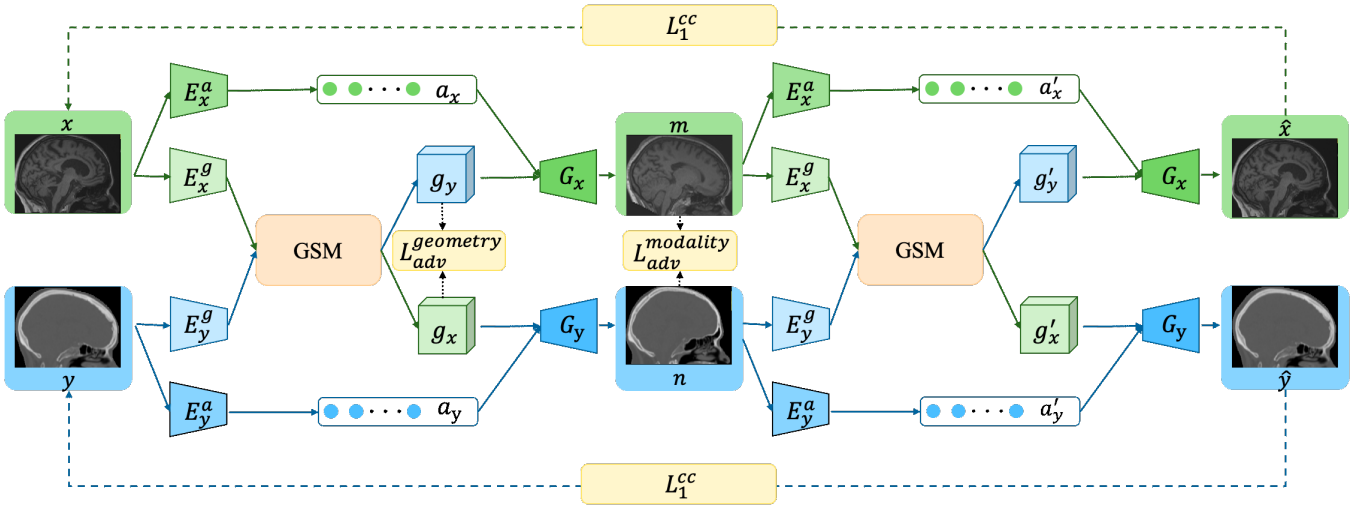

Fig. 1: Overview of our framework with geometry similarity guided feature disentanglement for MRI to CT synthesis.

of them have low intensity values, we propose a novel Geometry Similarity Module (GSM) to take context information into consideration.

2. METHOD

In this section, we first introduce the disentangled representation learning method, which can factorize MRI and CT images as modality-invariant geometry code and modality-specific appearance code. The overview of our proposed MRI to CT synthesis framework is presented in Fig.1.

2.1. Disentangle geometry and appearance representations

We denote the MRI slices and CT slices by $x \in \mathbb{R}^{H\times W\times 1}$ and $y \in \mathbb{R}^{H\times W\times 1}$ respectively. As shown in Fig.1, the framework is mainly composed of geometry encoders $\{E_x^g, E_y^g\}$, appearance encoders $\{E_x^a, E_y^a\}$, generators $\{G_x, G_y\}$, geometry similarity module $GSM(\cdot)$, modality discriminators $\{D_x, D_y\}$ and geometry discriminator D^g, where the $GSM(\cdot)$ will be introduced in detail in subsection 2.3. Since our goal is to achieve modal translation from x to y , here we take x as an example. For the input x, we can obtain its disentangled geometry code $g_x = GSM(E_x^g(x))$ and appearance code $a_x = E_x^a(x)$, which is set as 8-bit vector.

The disentangled representation learning of the brain CT and MR images is based on the fact that different modalities of the same patient share the same anatomical geometry structure while showing different appearance. Therefore, a shared domain-invariant space that retains the geometric information and a domain-specific appearance code for each modality can be explored to recover the underlying mapping between CT and MR images [12]. Inspired by [13], we employ weight-sharing and geometry discriminator for the implementation of disentangled representation learning. The weight-sharing we adopted includes two parts. The first part is sharing the weight between the last layer of E_x^g and E_y^g and the first layer of G_x and G_y. The second part is the shared GSM. However, weight sharing is not enough to ensure that the feature distributions of the two modalities tend to be consistent, so as to achieve the disentanglement of geometric structures. Therefore, generative adversarial learning [14] is introduced to further achieve the above goals. Specifically, a geometry discriminator D^g is proposed to distinguish the modality membership of the geometry representations g_x and g_y. Adversely, the geometry encoders E_x^g, $E_y^g$and shared GSM $GSM(\cdot)$ learn to generate geometry features that modality membership can not be distinguished by the geometry discriminator D^g. Based on the disentangled feature where the geometry space is shared among modalities and the appearance space encodes intra-modality variations, we can perform MRI to CT synthesis by combining a geometry representation from MRI with an appearance representation from an arbitrary image from CT modality. As illustrated in Fig.1, it can be expressed as $n = G_y(g_x, a_y)$.

2.2. Slice corresponding strategy

Disentangled representation learning is based on the premise that MRI and CT slices share a common geometric structure. However, since the slices of MRI and CT are not paired, in the actual training process, it could happen that the input MRI and CT slices are located in extremely different locations of the brain volume. Obviously, the premise of disentangled representation learning is not valid in this case, so we propose a Slice Corresponding (SC) strategy to facilitate reasonable disentanglement. The SC strategy can be presented as follows:

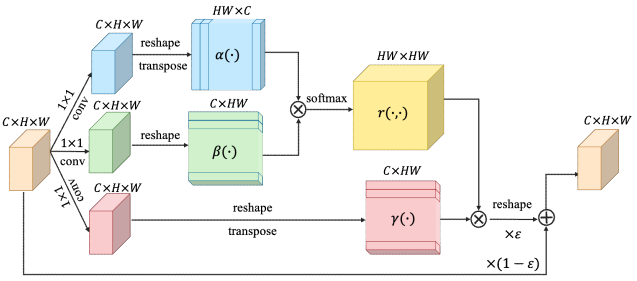

Fig. 2: The details of geometry similarity module (GSM)

$$x(i) = \begin{cases} [k * N(x)] + p, & if\ 3 \leq [k * N(x)] \leq N(x) - 3 \\ [k * N(x)], & otherwise \end{cases} \tag{1}$$

$$y(i) = \begin{cases} [k * N(y)] + q, & if\ 3 \leq [k * N(y)] \leq N(x) - 3 \\ [k * N(y)], & otherwise \end{cases} \tag{2}$$

where k represents a random value uniformly sampled from the range of $[0, 1]$, $N(\cdot)$ is a function to calculate the number of slices of a given volume data, $[\cdot]$ denote the rounding function, and p and q represent random integers sampled from the range of $[-3, 3]$.

2.3. Geometry similarity module

Rich and global contextual information is an important part of discriminant representation for pixel-level visual tasks. It was observed that it is a challenging task to distinguish cortical bones from air in MR images as both have low intensity values. To solve the challenge, we propose a Geometry Similarity Module (GSM) to explore abundant contextual information by capturing long-range dependencies.

The GSM structure is shown in Fig.2. Specifically, for a given feature map $F \in \mathbb{R}^{H \times W \times C}$, each pair of position (f_i, f_j) in F will be computed a correlation strength matrix $r(f_i, f_j) \in \mathbb{R}^{N \times N}$, where $N = H \times W$. Suggested by the non-local operation[15, 16], the r is defined as the following formula:

$$r(f_i, f_j) = \frac{exp(\alpha(f_i)^T \beta(f_j))}{\sum_{i=1}^{N} exp(\alpha(f_i)^T \beta(f_j))} \tag{3}$$

where the $\alpha(\cdot)$ and $\beta(\cdot)$ are composed of 1×1 convolution and reshape operation. The $r(f_i, f_j)$ indicates the i^{th} feature point's impact on j^{th} feature point. Intuitively, the more similar the two feature points, the stronger the strength of their association. In addition, we feed F to another 1×1 convolution layer to get a new feature map and reshape it to $\mathbb{R}^{C \times N}$, which could be expressed by $\gamma(\cdot)$. Unlike the attention mechanism, we did not directly add enhanced features to the original features. Instead, we compute a weighted sum of the enhanced features and the original features:

$$out = \varepsilon \sum_{i=1}^{N} r(f_i, f_j)\gamma(f_j) + (1 - \varepsilon)F \tag{4}$$

where hyper-parameter ε controls the relative importance of these two types of features and is empirically set to 0.2.

It can be inferred that the output feature at each point is a weighted sum of the relationship feature across all positions and the original feature. Therefore, our model takes contextual information into consideration and thus can be used to distinguish cortical bones from air during the CT synthesis process, even when they both have low intensity values in MR images.

2.4. Loss functions

In this section, all the loss functions used to facilitate network training are introduced in details. To achieve feature disentanglement, geometry adversarial loss is proposed to constrain the common modality-invariant geometry space of the two imaging modalities. We express this geometry adversarial loss as:

$$\begin{aligned} &L_{adv}^{geometry}(E_x^g, E_y^g, GSM, D^g) = \\ &\mathbb{E}_x[\frac{1}{2}logD^g(GSM(E_x^g(x))) + \frac{1}{2}log(1 - D^g(GSM(E_x^g(x))))]+ \\ &\mathbb{E}_y[\frac{1}{2}logD^g(GSM(E_y^g(y))) + \frac{1}{2}log(1 - D^g(GSM(E_y^g(y))))] \end{aligned} \tag{5}$$

Similarly, in order to constrain the generated MR and CT images as real as possible, we introduce the modality adversarial loss function $L_{adv}^{modality}$ where the discriminators D_x and D_y try to distinguish real images from synthetic images in each modality, while the generators G^x and G^y try to generate realistic images to fool the discriminators.

In addition to the adversarial loss, we also introduce two L_1 loss functions. The first one is cross-cycle consistency loss L_1^{cc}, which is a variant of cycle consistency loss proposed in [9] in order to adapt to the feature disentanglement. As illustrated in Fig.1, we formulate the cross-cycle consistency loss as:

$$\begin{aligned} &L_1^{cc}(G_x, G_y, E_x^g, E_y^g, E_x^a, E_y^a, GSM) = \\ &\mathbb{E}_{x,y}[\left\|G_x(GSM(E_y^g(n)), GSM(E_x^a(m))) - x\right\|_1 \\ &+ \left\|G_y(GSM(E_x^g(m)), GSM(E_y^a(n))) - y\right\|_1] \end{aligned} \tag{6}$$

where $m = G_x(g_y, a_x)$ and $n = G_y(g_x, a_y)$, respectively. The another L_1 loss function is about self-reconstruction. Take x as an example, from the view of perfect disentanglement, the obtained geometry representation g_x and appearance code a_x should enable to be re-rendered into original image through the generator G_x. To enhance the

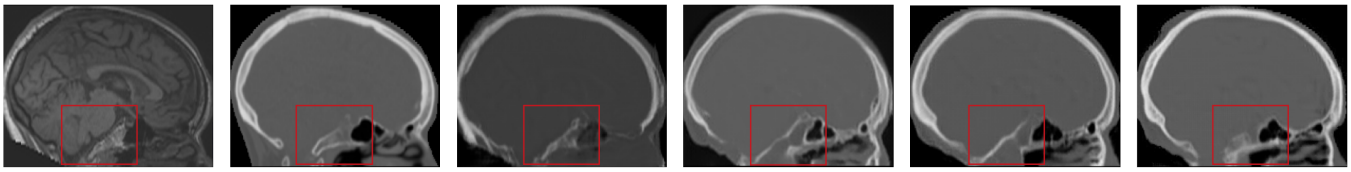

Fig. 3: Visualization of MRI to CT synthesis results by different methods

self-reconstruction ability, the following loss function is introduced as:

$$L_1^{rec}(G_x, G_y, E_x^g, E_y^g, E_x^a, E_y^a, GSM) = \mathbb{E}_x[\|G_x(GSM(E_x^g(x)), E_x^a(x)) - x\|_1] + \mathbb{E}_y[\|G_y(GSM(E_y^g(y)), E_y^a(y)) - y\|_1] \quad (7)$$

Overall, the loss functions in our framework could be expressed as following:

$$L_{D^g, D_x, D_y} = \lambda_{adv}^{geometry} L_{adv}^{geometry} + \lambda_{adv}^{modality} L_{adv}^{modality}$$
$$L_{E^g, E^a, G, GSM} = -L_{D^g, D_x, D_y} + \lambda_1^{rec} L_1^{rec} + \lambda_1^{cc} L_1^{cc} \quad (8)$$

where the hyper-parameters λs refers to the weight of each item.

3. EXPERIMENTS

3.1. Dataset

We evaluated our method on paired brain MRI and CT volumes from 48 patients as used in [11]. The database was split into a training set containing MR and CT volumes of 38 patients and a test set of 10 patients, in such a way that subjects do not appear in both datasets. Similar to [5, 17], the experiments were performed on 2D sagittal image slices. Before we get slices from the volume data, we first resample the voxel spacing of volume data to $1 \times 1 \times 1mm^3$. Each CT and MRI volume data includes approximately 160 sagittal slices with a size of about 200×160. The intensity ranges of CT and MRI are $[-1000, 3500]$ HU and $[0, 3500]$ respectively. To avoid overfitting, we apply a series of data augmentation methods including flipping horizontally, flipping vertically and random cropped to $[180, 140]$ to generate more training samples.

3.2. Experimental Results

We reproduced the conventional CycleGAN as a baseline. Additionally, we compared our method with the sc-CycleGAN[17] where they introduced a structure loss function to constrain the geometric structure of the synthetic CT to be consistent with the input MR images. In addition, we conduct ablation experiments to verify the effectiveness of the SC strategy and the GSM. We adopt commonly used mean absolute error (MAE), peak signal-to-noise ratio (PSNR), and structural similarity (SSIM) as quantitative evaluation metrics. It is worth to mention that the maximum value in PSNR and the dynamic range in SSIM are set to 4500 due to the intensity range of the CT data.

Table 1: MRI to CT synthesis accuracies for different synthesis methods

Method	MAE	PSNR	SSIM
CycleGAN	215.26	20.45	0.661
sc-CycleGAN	129.00	24.15	0.779
Ours, without (SC & GSM)	112.96	24.54	0.796
Ours, without GSM	107.72	24.85	0.804
Ours	103.82	25.34	0.827

Tab.1 shows that our proposed method yields significantly better results when compared to the baseline and the sc-CycleGAN in all evaluation metrics. At the same time, the results of the ablation study (the last three rows in Tab.1) proved that our proposed slice corresponding strategy and geometry similarity module could effectively improve the performance of our approach. The visual examples of synthetic CT obtained from the same test MRI when different methods were used are shown in Fig. 3. Without using GSM, it is difficult to distinguish cortical bones from air, leading to unrealistic synthesis results, as observed in Fig. 3 (red box). After incorporating GSM, our method generated more realistic synthesis results, proving the efficacy of the proposed module.

4. CONCLUSION

In summary, we presented a disentangled representation learning method from MR to CT synthesis using unpaired data. Unlike previous methods based on CycleGAN [5, 8, 10, 11], our method treated the MR to CT synthesis as a many-to-many image translation problem. Our results demonstrated that the present method achieved better or equivalent results than the state-of-the-art methods.

5. REFERENCES

[1] D Nie, X Cao, and et al., "Estimating ct image from mri data using 3d fully convolutional networks," in *Proc. DLMIA*. 2016, pp. 170–178, Springer.

[2] X Han, "Mr-based synthetic ct generation using a deep convolutional neural network," *Med. Phys.*, vol. 44, no. 4, pp. 1408–1419, 2017.

[3] N Burgos, F Guerreiro, and et al., "Iterative framework for the joint segmentation and ct synthesis of mr images: application to mri-only radiotherapy treatment planning," *Phys Med Biol*, vol. 62, pp. 4237–4253, 2017.

[4] D Nie, R Trullo, and et al., "Medical image synthesis with context-aware generative adversarial networks," in *Proc. MICCAI*, 2017, pp. 417–425.

[5] JM Wolterink, AM Dinkla, and et al., "Deep mr to ct synthesis using unpaired data," in *Proc. SASHIMI*, 2017, pp. 14–23.

[6] F Liu, H Jang, and et al., "Deep learning mr imaging-based attenuation correction for pet/mr imaging," *Radiology*, vol. 286, no. 2, pp. 676–684, 2017.

[7] A Chartsias, T Joyce, and et al., "Adversarial image synthesis for unpaired multi-modal cardiac data," in *Proc. SASHIMI*. Springer, 2017, pp. 3–13.

[8] Zizhao Zhang, Lin Yang, and Yefeng Zheng, "Translating and segmenting multimodal medical volumes with cycle-and shape-consistency generative adversarial network," in *Proc. CVPR*, 2018, pp. 9242–9251.

[9] J-Y amd Parl. T Zhu, P Isola, and AA Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. ICCV*. 2017, pp. 2242–2251, IEEE.

[10] Yongsheng Pan, Mingxia Liu, and et al., "Synthesizing missing pet from mri with cycle-consistent generative adversarial networks for alzheimer's disease diagnosis," in *Proc. MICCAI2018*. Springer, 2018, pp. 455–463.

[11] Guodong Zeng and Guoyan Zheng, "Hybrid generative adversarial networks for deep mr to ct synthesis using unpaired data," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 759–767.

[12] Junlin Yang, Nicha C Dvornek, Fan Zhang, Julius Chapiro, MingDe Lin, and James S Duncan, "Unsupervised domain adaptation via disentangled representations: Application to cross-modality liver segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 255–263.

[13] Hsin-Ying Lee, Hung-Yu Tseng, Qi Mao, Jia-Bin Huang, Yu-Ding Lu, Maneesh Singh, and Ming-Hsuan Yang, "Drit++: Diverse image-to-image translation via disentangled representations," *International Journal of Computer Vision*, pp. 1–16, 2020.

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, pp. 2672–2680, 2014.

[15] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.

[16] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu, "Dual attention network for scene segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3146–3154.

[17] Heran Yang, Jian Sun, Aaron Carass, Can Zhao, Junghoon Lee, Zongben Xu, and Jerry Prince, "Unpaired brain mr-to-ct synthesis using a structure-constrained cyclegan," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pp. 174–182. Springer, 2018.