



# Towards cross-modal organ translation and segmentation: A cycle- and shape-consistent generative adversarial network<sup>☆</sup>



Jinzheng Cai<sup>a,1</sup>, Zizhao Zhang<sup>b,1</sup>, Lei Cui<sup>c</sup>, Yefeng Zheng<sup>d,2</sup>, Lin Yang<sup>a,\*</sup>

<sup>a</sup>J. Crayton Pruitt Family Department of Biomedical Engineering, University of Florida, Gainesville, FL, 32611, USA

<sup>b</sup>Department of Computer Information and Science Engineering, University of Florida, Gainesville, FL, 32611, USA

<sup>c</sup>Department of Information Science and Technology, Northwest University, Xi'an 710127, China

<sup>d</sup>Medical Imaging Technologies, Siemens Healthcare, Princeton, NJ, 08540, USA

## ARTICLE INFO

### Article history:

Received 29 May 2018

Revised 12 December 2018

Accepted 14 December 2018

Available online 19 December 2018

### Keywords:

Medical image synthesis

Generative adversarial network (GAN)

Computed tomography (CT)

Magnetic resonance imaging (MRI)

Mammography X-ray

Organ segmentation

## ABSTRACT

Synthesized medical images have several important applications. For instance, they can be used as an intermedium in cross-modality image registration or used as augmented training samples to boost the generalization capability of a classifier. In this work, we propose a generic cross-modality synthesis approach with the following targets: 1) synthesizing realistic looking 2D/3D images without needing paired training data, 2) ensuring consistent anatomical structures, which could be changed by geometric distortion in cross-modality synthesis and 3) more importantly, improving volume segmentation by using synthetic data for modalities with limited training samples. We show that these goals can be achieved with an end-to-end 2D/3D convolutional neural network (CNN) composed of mutually-beneficial generators and segmentors for image synthesis and segmentation tasks. The generators are trained with an adversarial loss, a cycle-consistency loss, and also a shape-consistency loss (supervised by segmentors) to reduce the geometric distortion. From the segmentation view, the segmentors are boosted by synthetic data from generators in an online manner. Generators and segmentors prompt each other alternatively in an end-to-end training fashion. We validate our proposed method on three datasets, including cardiovascular CT and magnetic resonance imaging (MRI), abdominal CT and MRI, and mammography X-rays from different data domains, showing both tasks are beneficial to each other and coupling these two tasks results in better performance than solving them exclusively.

© 2018 Published by Elsevier B.V.

<sup>☆</sup> **Conflicts of Interest Statement:** We wish to draw the attention of the Editor to the following facts which may be considered as potential conflicts of interest and to significant financial contributions to this work. We confirm that the manuscript has been read and approved by all named authors and that there are no other persons who satisfied the criteria for authorship but are not listed. We further confirm that the order of authors listed in the manuscript has been approved by all of us. We confirm that we have given due consideration to the protection of intellectual property associated with this work and that there are no impediments to publication, including the timing of publication, with respect to intellectual property. In so doing we confirm that we have followed the regulations of our institutions concerning intellectual property. We understand that the Corresponding Author is the sole contact for the Editorial process (including Editorial Manager and direct communications with the office). He/she is responsible for communicating with the other authors about progress, submissions of revisions and final approval of proofs. We confirm that we have provided a current, correct email address which is accessible by the Corresponding Author and which has been configured to accept email from [lin.yang@bme.ufl.edu](mailto:lin.yang@bme.ufl.edu)

\* Corresponding author.

E-mail addresses: [jimmycail@ufl.edu](mailto:jimmycail@ufl.edu) (J. Cai), [lin.yang@bme.ufl.edu](mailto:lin.yang@bme.ufl.edu) (L. Yang).

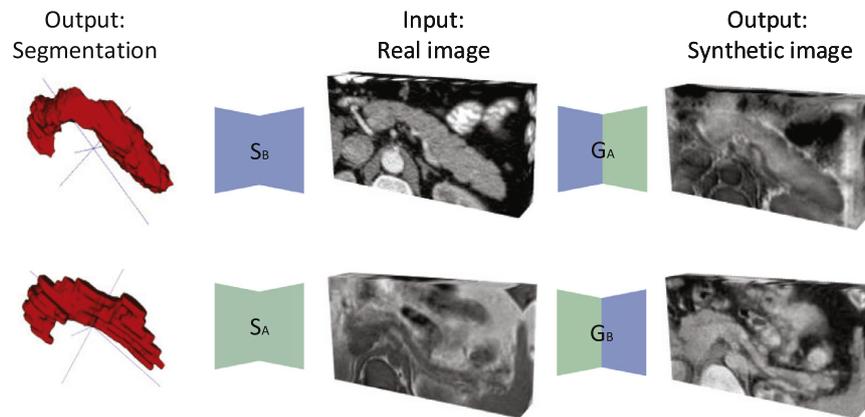
<sup>1</sup> Equal contribution.

<sup>2</sup> Present address: Medical AI, Tencent, Shenzhen 518052, China.

## 1. Introduction

In current clinical practice, multiple imaging modalities may be available for disease diagnosis and surgical planning (Cao et al., 2017; Chen et al., 2017). For a specific patient group, a certain imaging modality might be more popular than others. Due to the proliferation of multiple imaging modalities, there is a strong clinical need to develop a cross-modality image transfer analysis system to assist clinical treatment. For example, synthesized computed tomography (CT) data can provide X-ray attenuation map for radiation therapy planning (Burgos et al., 2015).

Neural networks have been widely used for medical image analysis, such as detection (Roth et al., 2014), segmentation (Ronneberger et al., 2015; Roth et al., 2018), and classification (Zhang et al., 2017a, b). Such methods are often generic and can be extended from one imaging modality to the other by re-training on the target imaging modality. However, a sufficient number of representative training images are required to achieve enough robustness. In practice, it is often difficult to collect enough training images, especially for a new imaging modality not well established



**Fig. 1.** Our method learns two parallel sets of generators  $G_{A/B}$  and segmentors  $S_{A/B}$  for two modalities A and B to translate and segment holistic 3D volumes. Here we illustrate using CT and MRI pancreatic 3D images. Best viewed in color.

in clinical practice yet. Synthesized data are often used to as supplementary training data in hope that they can boost the generalization capability of a trained neural network. This paper presents a novel method to address the above-mentioned two demanding tasks simultaneously (Fig. 1). The first is cross-modality (domain) translation and the second is improving segmentation models by making use of synthesized data.

To synthesize medical images, recent advances (Nie et al., 2018; Costa et al., 2018) have used generative adversarial networks (GANs) (Goodfellow et al., 2014) to formulate it as an image-to-image translation task. These methods require pixel-to-pixel correspondence between two domain data to build direct cross-modality reconstruction. However, in a more common scenario, multimodal medical images are in 3D and do not have cross-modality paired data. A method to learn from unpaired data is the more general purpose. Furthermore, tomography structures (e.g., shape), in medical images/volumes, contain diagnostic information. Keeping their invariance in translation is critical. However, when using GANs without paired data, due to the lack of direct reconstruction, relying on discriminators to guarantee this requirement is not enough as we explain later.

It is an active research area by using synthetic data to overcome the insufficiency of labeled data in CNN training. In the medical image domain, people are interested in learning unsupervised translation between different modalities (Kamnitsas et al., 2017; Zbontar et al., 2018), so as to transfer existing labeled data from other modalities. However, the effectiveness of synthetic data heavily depends on the distribution gap between real and synthetic data. A possible solution to reduce such gap is by matching their distributions through GANs (Shrivastava et al., 2017; Bousmalis et al., 2017).

In this paper, we present a general-purpose method to realize both medical volume translation as well as segmentation. In brief, given two sets of unpaired data in two modalities, we simultaneously learn generators for cross-domain volume-to-volume translation and stronger segmentors by taking advantage of synthetic data translated from another domain. Our method is composed of several 3D/2D CNNs. From the generator learning view, we propose to train adversarial networks with cycle-consistency (Zhu et al., 2017a) to solve the problem of data without correspondence. We then propose a novel shape-consistency scheme to guarantee the shape invariance of synthetic images, which is supported by another CNN, namely segmentor. From the segmentor learning view, segmentors directly take advantage of generators by using synthetic data to boost the segmentation performance in an online fashion. Both generator and segmentor can take benefits from another in end-to-end training with one optimization objective.

We conduct extensive experiments with cardiac 3D image in MRI and CT, pancreatic abdomen scans in MRI and CT, and mammography X-rays from two independent domains. Comprehensive experimental results suggest that the proposed method achieves realistic modality translation and significantly improves the performance of segmentation models with a variety of experimented architectures. We demonstrate that using synthetic data as an isolated offline data augmentation process underperforms our end-to-end online approach. On the image segmentation task, blindly using synthetic data with a small number of real data can even distract the optimization when trained in the offline fashion. However, our method does not have this problem and leads to consistent improvement.

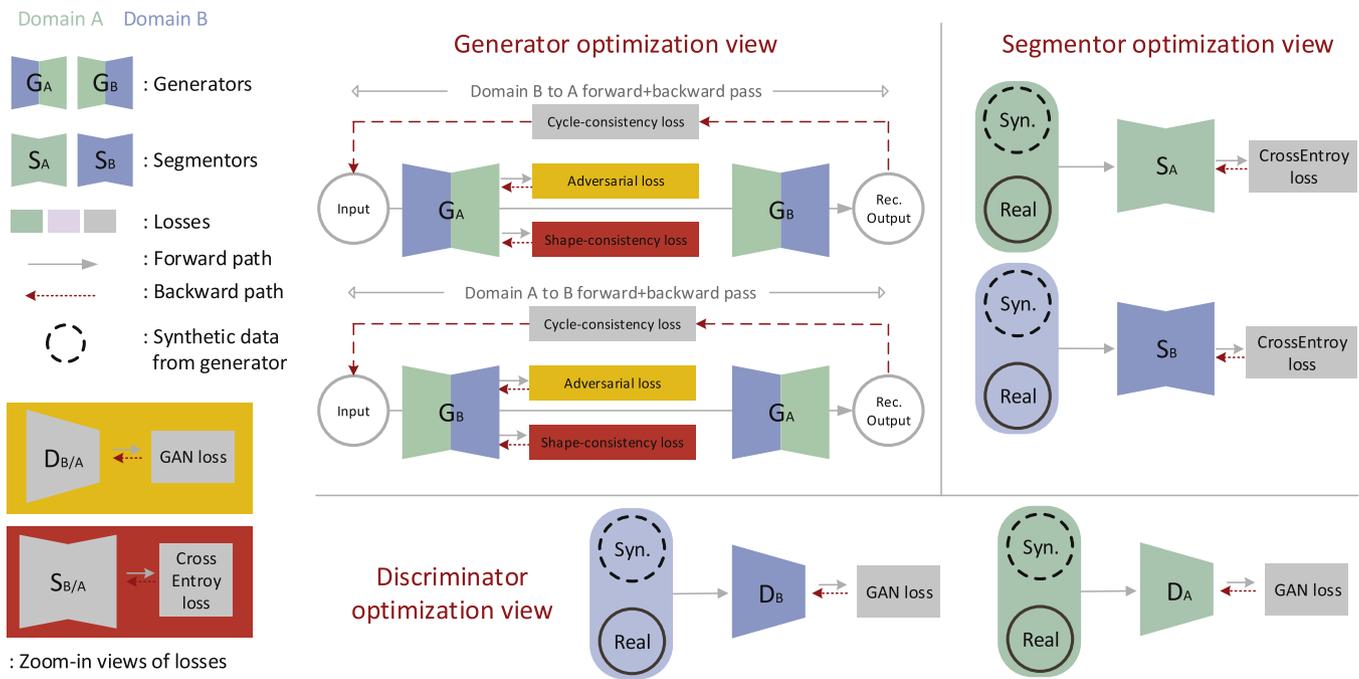
A preliminary version of this paper has been published in a conference paper (Zhang et al., 2018b), which is evaluated on cardiac 3D images. In this paper, we have made significant extensions to generalize our methods on multiple major 2D/3D imaging domains, aiming to provide a strong and comprehensive baseline for relevant research. To be specific,

1. This paper shows extensive experiments to validate the proposed method, including CT and MRI translation for pancreas segmentation and domain adaptation for mammogram X-rays for breast lesion segmentation.
2. This paper validates our method using a variety of advanced segmentation networks, including PSP-Net (Zhao et al., 2017), U-Net (Ronneberger et al., 2015), and Refine-Net (Lin et al., 2017) and show that our method performs generally well and consistently boosts medical 2D/3D image segmentation performance.
3. This paper systemically analyzes the effect of synthetic data on segmentation, with the goal to investigate the limitation of synthetic data, and inspire new scopes.

## 2. Related work

There are two demanding goals in medical image synthesis. The first is synthesizing realistic cross-modality images (Huang et al., 2018; Nie et al., 2018), and the second is to use synthetic data from other modalities with sufficient labeled data to help classification tasks (e.g., domain adaptation (Kamnitsas et al., 2017)).

In computer vision, recent image-to-image translation is formulated as a pixel-to-pixel mapping using encoder-decoder CNNs (Isola et al., 2017; Zhu et al., 2017a; Kim et al., 2017; Liu et al., 2017; Gong et al., 2017). Several studies have explored cross-modality translation for medical images, using probabilistic



**Fig. 2.** The illustration of our method from the generator optimization view (top-left), the segmentor optimization view (top-right), and the discriminator optimization view (bottom). The left panel illustrates each architecture components. Domain A and Domain B are illustrated using two colors. **Generator view:** Two generators learn cross-domain translation between domains A and B, which are supervised by a cycle-consistency loss and a combination of an adversarial loss and a shape-consistency loss (supported by segmentors), respectively. **Segmentor view:** Segmentors are trained by real data and extra synthetic data translated from domain-specific generators. **Discriminator view:** Discriminators are trained to distinguish real and synthetic data. The three optimization views are trained jointly and end-to-end. Best viewed in color.

generative model (Cordier et al., 2016), sparse coding (Huang et al., 2018; Vemulapalli et al., 2015), GANs (Nie et al., 2018; Osokin et al., 2017), CNN (Van Nguyen et al., 2015), etc. GANs have attracted wide interests in helping addressing such tasks to generate high-quality, less blurry results (Goodfellow et al., 2014; Arjovsky et al., 2017; Berthelot et al., 2017; Zhang et al., 2018a). More recent studies apply pixel-to-pixel GANs for brain MRI to CT image translation (Nie et al., 2018; Kamnitsas et al., 2017) and retinal vessel annotation to image translation (Costa et al., 2018). However, these methods presume targeting images have paired cross-domain data. Learning from unpaired cross-domain data is an attractive yet not well explored problem (Vemulapalli et al., 2015; Liu and Tuzel, 2016).

Synthesizing medical data to overcome insufficient labeled data attracted wide interests recently (Shrivastava et al., 2017; Zhao et al., 2018; Iglesias et al., 2013; Huo et al., 2018; Peng et al., 2018). Due to the diversity of medical modalities, learning an unsupervised translation between modalities is a promising direction (Costa et al., 2018). Kamnitsas et al. (2017) demonstrates the benefits on brain (MRI and CT) images, by using synthetic data as augmented training data to help lesion segmentation.

Apart from synthesizing data, several studies (Kohl et al., 2017; Luc et al., 2016; Yang et al., 2017; Xue et al., 2017) use adversarial learning as an extra supervision on the segmentation or detection networks. The adversarial loss plays a role of constraining the prediction to be close to the distribution of groundtruth. However, such strategy is a refinement process, so it is less likely to remedy the cost of data insufficiency.

### 3. Proposed method

#### 3.1. Multi-modal volume segmentation

There are a remarkable number of studies on medical image segmentation which is a key task in medical image analysis. Con-

volutional neural network (CNN) in nature needs to be fed with large data. However, medical data annotation is expensive. This paper mainly explore the possibility of augmenting realistic synthetic data to help training segmentation networks.

Our method can be understood in two views: the generator view and the segmentor view. From the segmentor view (Fig. 2), the goal is quite straightforward. We expect to train dual-modality segmentation networks  $S_A$  and  $S_B$  (namely segmentors) for domain A and B data, respectively. The synthetic volumes, provided by the generators (discussed below), provide augmented training dual-domain data to help improve the segmentors. During training,  $S_A$  and  $S_B$  receive both real data and synthetic data that are generated by generators online (see Fig. 2). Our method is model-agnostic. We show that our method successfully improves several popular segmentation architectures in experiments.

Note that the most straightforward way to use synthetic data is fusing them with real data and then train a segmentation CNN. We denote this as an ad-hoc offline data augmentation approach. Compared with it, our method implicitly performs data augmentation in an online manner. Formulated in our optimization objective, our method can use synthetic data adaptively, which thereby offers more stable training and thereby better performance than the offline approach.

In the following sections, we introduce image-to-image translation on medical 2D/3D images with our proposed method. Meanwhile, the segmentors assist the generators to guarantee shape-consistency.

#### 3.2. Image-to-image translation for unpaired data

Generative adversarial networks (GANs) have been widely used for image translation in the applications that need pixel-to-pixel mapping, such as image style transfer (Zhang et al., 2016). ConditionalGAN (Isola et al., 2017) shows a strategy to learn such translation mapping with a conditional setting to capture structure

information. However, it needs paired cross-domain images for the pixel-wise reconstruction loss. For some types of translation tasks, acquiring paired training data from two domains is difficult or even impossible. Recently, CycleGAN (Zhu et al., 2017a) and other similar methods (Kim et al., 2017; Yi et al., 2017) are proposed to generalize the ConditionalGAN to address this issue. Here we use CycleGAN to illustrate the key idea.

Given a set of unpaired data from two domains,  $A$  and  $B$ , CycleGAN learns two mappings,  $G_B: A \rightarrow B$  and  $G_A: B \rightarrow A$ , with two generators  $G_A$  and  $G_B$ , at the same time. To bypass the infeasibility of pixel-wise reconstruction with paired data, i.e.  $G_B(A) \approx B$  or  $G_A(B) \approx A$ , CycleGAN introduces an effective cycle-consistency loss for  $G_A(G_B(A)) \approx A$  and  $G_B(G_A(B)) \approx B$ . The idea is that the generated target domain data is able to return back to the exact data in the source domain it is generated from. To guarantee the fidelity of fake data  $G_B(A)$  and  $G_A(B)$ , CycleGAN uses two discriminators  $D_A$  and  $D_B$  to distinguish real or synthetic data and thereby encourage generators to synthesize realistic data (Goodfellow et al., 2014).

### 3.3. Problems in unpaired volume-to-volume translation

Lacking supervision with a direct reconstruction error between  $G_B(A)$  and  $B$  or  $G_A(B)$  and  $A$  brings some uncertainties and difficulties towards the desired outputs for more specified tasks. And it is even more challenging when training on 3D CNNs.

To be specific, cycle-consistency has an intrinsic ambiguity with respect to geometric transformations. For example, suppose generation functions,  $G_A$  and  $G_B$ , are cycle consistent, e.g.,  $G_A(G_B(A)) = A$ . Let  $T$  be a bijective geometric transformation (e.g., translation, rotation, scaling, or even nonrigid transformation) with inverse transformation  $T^{-1}$ .

It is easy to show that,

$$G'_A = G_A \circ T \text{ and } G'_B = G_B \circ T^{-1} \quad (1)$$

are also cycle consistent. Here,  $\circ$  denotes the concatenation operation of two transformations. That means, using CycleGAN, when an image is translated from one domain to the other it can be geometrically distorted. And the distortion can be recovered when it is translated back to the original domain without provoking any penalty in data fidelity cost. From the discriminator perspective, geometric transformation does not change the realness of synthesized images since the shape of training data is arbitrary.

Such problem can destroy anatomical structures in synthetic medical volumes, which, however, has not been addressed by existing methods. Our solution is to add a shape-consistency condition to prevent such failure (see below).

### 3.4. Volume-to-volume cycle-consistency

To solve the task of learning generators with unpaired volumes from two domains,  $A$  and  $B$ , we adopt the idea of the cycle-consistency loss (described above) for generators  $G_A$  and  $G_B$  to force the reconstructed synthetic sample  $G_A(G_B(x_A))$  and  $G_B(G_A(x_B))$  to be identical to their inputs  $x_A$  and  $x_B$ :

$$\mathcal{L}_{cyc}(G_A, G_B) = \mathbb{E}_{x_A \sim p_d(x_A)} [ \|G_A(G_B(x_A)) - x_A\|_1 ] + \mathbb{E}_{x_B \sim p_d(x_B)} [ \|G_B(G_A(x_B)) - x_B\|_1 ], \quad (2)$$

where  $x_A$  is a sample from domain  $A$  and  $x_B$  is from domain  $B$ .  $p_d(\cdot)$  denotes the data distribution at a certain domain.  $\mathcal{L}_{cyc}$  uses the L1 loss over all voxels (or pixels), which shows better visual results than the L2 loss.

### 3.5. Volume-to-volume shape-consistency

To solve the intrinsic ambiguity with respect to geometric transformations in cycle-consistency as we pointed out above, our

method introduces two auxiliary mappings using the segmentors defined above, i.e.,  $S_A: A \rightarrow Y$  and  $S_B: B \rightarrow Y$ , to constrain the geometric invariance of synthetic data. They map the translated data from respective domain generators into a shared shape space  $Y$  (i.e. a semantic label space) and compute pixel-wise semantic ownership. The two mappings are represented by two CNNs, namely segmentors as described above for segmentation.

Given a real volume  $x_A$  and its translation  $G_B(x_A)$ . To encourage them to have the same shape, the key idea is to regularize  $G_B$  such that the difference (Diff) of them in the space  $Y$  to be small. We can directly optimize the following Diff to achieve this goal

$$\min_{G_B} \text{Diff}(S_B(G_B(x_A)), y_A) \quad (3)$$

where  $y_A$  denotes the ground truth shape representation of the real sample volume  $x_A$ . We specify Diff as a standard multi-class cross-entropy loss for semantic segmentation (use  $S_B$  as the sample)

$$\text{Diff}(S_B(G_B(x_A)), y_A) = -\frac{1}{N_A} \sum_i y_A^i \log(S_B(G_B(x_A))_i), \quad (4)$$

and define our shape-consistency loss as

$$\begin{aligned} \mathcal{L}_{shape}(S_A, S_B, G_A, G_B) &= \mathbb{E}_{x_B \sim p_d(x_B)} \left[ -\frac{1}{N_B} \sum_i y_B^i \log(S_A(G_A(x_B))_i) \right] \\ &+ \mathbb{E}_{x_A \sim p_d(x_A)} \left[ -\frac{1}{N_A} \sum_i y_A^i \log(S_B(G_B(x_A))_i) \right], \end{aligned} \quad (5)$$

where  $y_A^i, y_B^i \in \{0, 1, \dots, C\}$  represent one voxel with one out of  $C$  classes.  $N_A$  and  $N_B$  are the total numbers of voxels in volume  $x_A$  and  $x_B$ , respectively. This objective optimizes  $G_{A/B}$  (refers to  $G_A$  and  $G_B$  in brevity) and keeps  $S_{A/B}$  fixed. Optimizing  $S_{A/B}$  is just feeding in real and synthetic data together and updating them. Recall that by feeding synthetic data here, we successfully transfer diverse data with sufficient segmentation annotation (see Fig. 2 right) to the target domain and thereby improve domain-specific segmentors, as described in Section 3.1.

**Regularization** Shape-consistency provides a level of regularization on generators. Recall that different from ConditionalGAN, since we have no paired data, the only supervision for  $G_A(x_B)$  and  $G_B(x_A)$  is the adversarial loss, which is not sufficient to preserve all types of information in synthetic images, such as the annotation correctness. Shrivastava et al. (2017) introduces a self-regularization loss between an input image and an output image to force the annotations to be preserved. Our shape-consistency performs a similar role to preserve pixel-wise semantic label ownership, as a way to regularize the generators and guarantee the anatomical structure invariance in medical volumes.

### 3.6. Objective

Given the definitions of cycle-consistency and shape-consistency losses above, we define our full objective for optimizing  $G_A$  and  $G_B$  as:

$$\begin{aligned} \mathcal{L}(G_A, G_B, D_A, D_B, S_A, S_B) &= \mathcal{L}_{GAN}(G_A, D_A) \\ &+ \mathcal{L}_{GAN}(G_B, D_B) \\ &+ \lambda \mathcal{L}_{cyc}(G_A, G_B) \\ &+ \gamma \mathcal{L}_{shape}(S_A, S_B, G_A, G_B) \end{aligned} \quad (6)$$

The adversarial loss  $\mathcal{L}_{GAN}$  is defined as

$$\begin{aligned} \mathcal{L}_{GAN}(G_B, D_B) &= \mathbb{E}_{x_B \sim p_d(x_B)} [ \|D_B(x_B)\|_2 ] \\ &+ \mathbb{E}_{x_A \sim p_d(x_A)} [ \|D_B(G_B(x_A)) - \mathbf{1}\|_2 ] \end{aligned} \quad (7)$$

where  $\mathbf{1}$  denotes a matrix that has the same size with the output of  $D_{A/B}$  but has all values equal to 1. The definition for  $\mathcal{L}_{GAN}(G_A, D_A)$  is identical. This objective is trained in an adversarial manner,  $G_B$  tries to minimize this objective and  $D_B$  tries to maximize it. They resemble the usage of the methods presented by Zhu et al. (2017a) and Isola et al. (2017). We follow the Least Squares GAN (LSGAN) (Mao et al., 2017) loss to replace the original cross-entropy loss in GANs with the simple mean squared loss.  $\lambda$  is set to 10 and  $\gamma$  is set to 1 during training. To optimize the overall networks, we update them alternatively: optimizing generators  $G_{A/B}$  with segmentors  $S_{A/B}$  and discriminators  $D_{A/B}$  fixed and then optimizing  $S_{A/B}$  and  $D_{A/B}$  (they are independent), respectively, with  $G_{A/B}$  fixed.

The generators and segmentors are mutually beneficial, because to make the full objective optimized, the generators have to generate synthetic data with lower shape-consistency loss, which, from another angle, indicates lower segmentation losses over synthetic training data.

#### 4. Network architecture and details

This section discusses necessary architecture and training details for generating high-quality 3D images.

##### 4.1. Architecture

Training deep networks end-to-end on 3D images is much more difficult (from optimization and memory aspects) than 2D images. Instead of using 2.5D (Roth et al., 2014) or sub-volumes (Kamnitsas et al., 2017), our method directly deals with holistic volumes. Our design trades-off network size and maximizes its effectiveness. There are several keys of network designs in order to achieve visually better results. The architecture of our method is composed by 3D fully convolutional layers with instance normalization (Ulyanov et al., 2016) (performs better than batch normalization (Ioffe and Szegedy, 2015)) and ReLU for generators or LeakyReLU for discriminators. CycleGAN originally designs generators with multiple residual blocks (He et al., 2016). Differently, in our generators, we make several critical modifications with justifications.

First, we find that using both bottom and top layer representations are critical to maintain the anatomical structures in medical images. We use long-range skip-connection in U-net (Ronneberger et al., 2015) as it achieves much faster convergence and locally smooth results. ConditionalGAN also uses U-net generators, but we do not downsample feature maps as greedily as it does. We apply 3 times downsampling with stride-2  $3 \times 3 \times 3$  convolutions totally, so the maximum downsampling rate is 8. The upsampling part is symmetric. Two sequential convolutions are used for each resolution, as it performs better than using one. Second, we replace transpose-convolutions to stride 2 nearest upsampling followed by a  $3 \times 3 \times 3$  convolution to realize upsampling as well as channel changes. It is also observed in Odena et al. (2016) that transpose-convolution can cause checkerboard artifacts due to the uneven overlapping of convolutional kernels. Actually, this effect is even severer for 3D transpose-convolutions as one pixel will be covered by  $2^3$  overlapping kernels (results in 8 times uneven overlapping). Fig. 3 compares the results with CycleGAN, demonstrating that our method can obtain significantly better visual quality.<sup>3</sup>

For discriminators, we adopt the PatchGAN proposed by Shrivastava et al. (2017) to classify whether an overlapping sub-

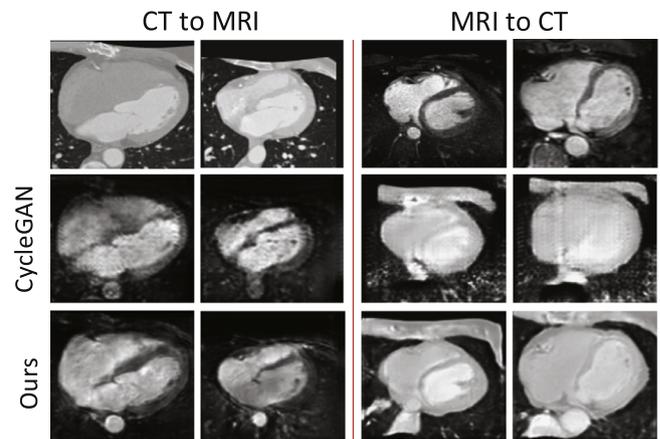


Fig. 3. Example outputs on 2D slides of 3D cardiovascular CT and MRI images of the results using 3D CycleGAN (second row) and ours (third row). The first row is the input samples. The original results of CycleGAN have severe artifacts, checkerboard effects, and missing anatomies (e.g., descending aorta and spine), while our method overcomes these issues and achieves significantly better quality.

volume is real or fake, rather than to classify the whole volume. Such approach limits discriminators to use unexpected information from arbitrary volume locations to make decisions. Details of the network configurations of the generators and discriminators are presented in Table 1.

For segmentors, we use U-Net (Ronneberger et al., 2015), Refine-Net (Lin et al., 2017), and PSP-Net (Zhao et al., 2017) to evaluate the performance improvement of our method, with the goal to demonstrate that our method is robust to different segmentor frameworks.

To be specific, for U-Net, 3 times symmetric downsampling and upsampling are performed by stride 2 max-pooling and nearest upsampling. For each resolution, we use two sequential convolutions. For Refine-Net, we keep its downsampling branch the same configuration as the U-Net and accordingly implement one Refine-Net unit at each resolution scale in the upsampling branch. For PSP-Net, to conduct equal comparison, we use the same downsampling branch as U-Net and Refine-Net. The PSP-Net aggregates feature maps of three resolutions in its pyramid pooling module for image segmentation. To make these segmentors seamlessly take in CT and MRI scans, we implement them with  $3 \times 3 \times 3$  convolutions so as to directly process images in 3D.

The Refine-Net upgrades U-Net by replacing U-Net's long-range skip connection with long-range residual connection presenting a more efficient gradient back-propagation through the network all the way to early low-level layers. Differently, PSP-Net simplifies the upsampling branch of U-Net by replacing its deconvolution-convolution units with direct bilinear interpolation resulting a much lighter model size. Thus, the segmentors ranked by their model capacities are PSP-Net, U-Net, and Refine-Net, in a descending order. We use such knowledge to investigate the correlation between the number of synthetic images and the segmentor capacity.

##### 4.2. Training details

We use the Adam solver (Kingma and Ba, 2014) for segmentors with a learning rate of  $2e-4$  and closely follow the settings in CycleGAN to train generators with discriminators. In the next section, for the purpose of fast experimenting, we choose to pre-train the  $G_{A/B}$  and  $D_{A/B}$  separately first and then train the whole network jointly. We hypothesized that fine-tuning generators and segmentors first is supposed to have better performance because they only affect each other after they have the sense of reasonable outputs.

<sup>3</sup> We have experimented many different configurations of generators and discriminators. All trials did not achieve desired visual results compared with our configuration.

**Table 1**

Network configurations of generators and discriminators implemented in the proposed CycleGAN. We denote network components *LeakyReLU* and *InstanceNormalization3d* as *LReLU* and *IN3d*, respectively. We denote the scale-2 nearest upsampling operation, and concatenation operation as *Upsample2* and *Concat*, respectively. We present convolutional layers and the hyperparameters in the form of *Conv3d*(*in channel*, *out channel*, *kernel size*, *stride*, *padding*). ‘|’ denotes the skip connection of the generator. Size of inputs are shown as (batch\_size × depth × width × height × channel).

$G_A$	$D_A$
Inputs( $b \times 80 \times 128 \times 128 \times 1$ )	Inputs( $b \times 80 \times 128 \times 128 \times 1$ )
LReLU(IN3d(Conv3d(1, 64, 3, 2, 1)))	LReLU(Conv3d(1, 64, 4, 2, 1))
IN3d(Conv3d(64, 64, 3, 1, 1))	LReLU(IN3d(Conv3d(64, 128, 4, 2, 1)))
LReLU(IN3d(Conv3d(64, 256, 3, 2, 1)))	LReLU(IN3d(Conv3d(128, 256, 4, 2, 1)))
IN3d(Conv3d(256, 256, 3, 1, 1))	LReLU(IN3d(Conv3d(256, 512, 4, 1, 1)))
LReLU(IN3d(Conv3d(256, 256, 3, 2, 1)))	Sigmoid(Conv3d(512, 1, 4, 1, 1))
IN3d(Conv3d(256, 256, 3, 1, 1))	
Upsample2(LReLU())	
LReLU(IN3d(Conv3d(256, 256, 3, 2, 1)))	
IN3d(Conv3d(256, 256, 3, 1, 1))	
Upsample2(LReLU(Concat()))	
LReLU(IN3d(Conv3d(512, 64, 3, 2, 1)))	
IN3d(Conv3d(64, 64, 3, 1, 1))	
Upsample2(LReLU(Concat()))	
LReLU(IN3d(Conv3d(128, 128, 3, 2, 1)))	
Tanh(Conv3d(128, 1, 3, 1, 1))	

Nevertheless, we observed that training all from scratch can also obtain similar results. It demonstrates the effectiveness to couple both tasks in an end-to-end network and make them converge harmonically. We pre-train segmentors for 100 epochs and generators for 60 epochs. After jointly training for 50 epochs, we decrease the learning rates for both generators and segmentors steadily for 50 epochs till 0. We found that if the learning rate decreases to a certain small value, the synthetic images turn to show clear artifacts and the segmentors tend to overfit. We apply early stop when the segmentation loss no longer decreases for about 5 epochs (usually takes 40 epochs to reach a desired point). In training, the number of training data in two domains can be different. We go through all data in the domain with a larger amount as one epoch.

## 5. Experiments

In this section, the effectiveness of the proposed method is demonstrated with applications in multiple contexts. First, it translates images between CT and MRI for non-rigid organ segmentation, e.g., cardiac in chest and pancreas in abdomen. It is then applied to translate images between mammography X-rays from two significantly different domains. For all of these applications, our method is targeting on simultaneously producing high-quality synthesized images and improving target object (e.g., organ and mass) segmentation.

### 5.1. Materials

**Cardiac datasets** We introduce a 3D cardiovascular image dataset. Heart is a perfect example of the difficulty in getting paired cross-modality data as it is a nonrigid organ and it keeps beating. Even if there are CT and MRI scans from the same patient, they cannot be perfectly aligned. Then we evaluate the two tasks we addressed in our method, i.e., volume segmentation and synthesis, both qualitatively and quantitatively with our proposed auxiliary evaluation metrics.

We collected 4354 contrasted cardiac CT scans from patients with various cardiovascular diseases (2–3 volumes per patients). In addition, we collected 142 cardiac MRI scans with a new compressed sensing scanning protocol. This true 3D MRI scan with isotropic voxel size is a new imaging modality, only available in handful top hospitals. We crop  $86 \times 112 \times 112$  volumes around the

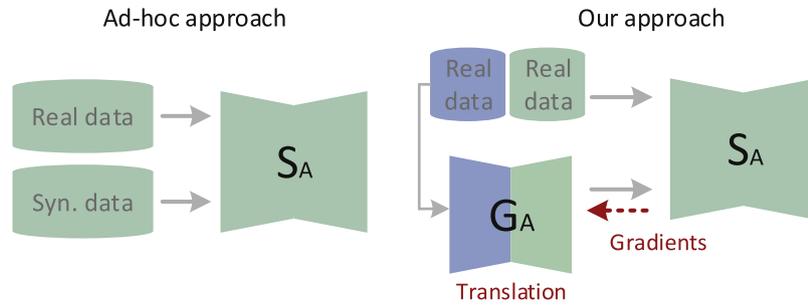
heart center. The endocardium of all four cardiac chambers is annotated. The left ventricle epicardium is annotated too, resulting in five anatomical regions.

We denote CT as domain  $A$  data and MRI as domain  $B$ . We organize the dataset in two sets  $S_1$  and  $S_2$ . For  $S_1$ , we randomly select 142 CT volumes from all CT images to match the number of MRI volumes. For both modalities, 50% data is used as training and validation and the rest 50% as testing data. For  $S_2$ , we use all the rest 4212 CT volumes as an extra augmentation dataset, which is used to generate synthetic MRI volumes for segmentation. We fix the testing data in  $S_1$  for all experiments.

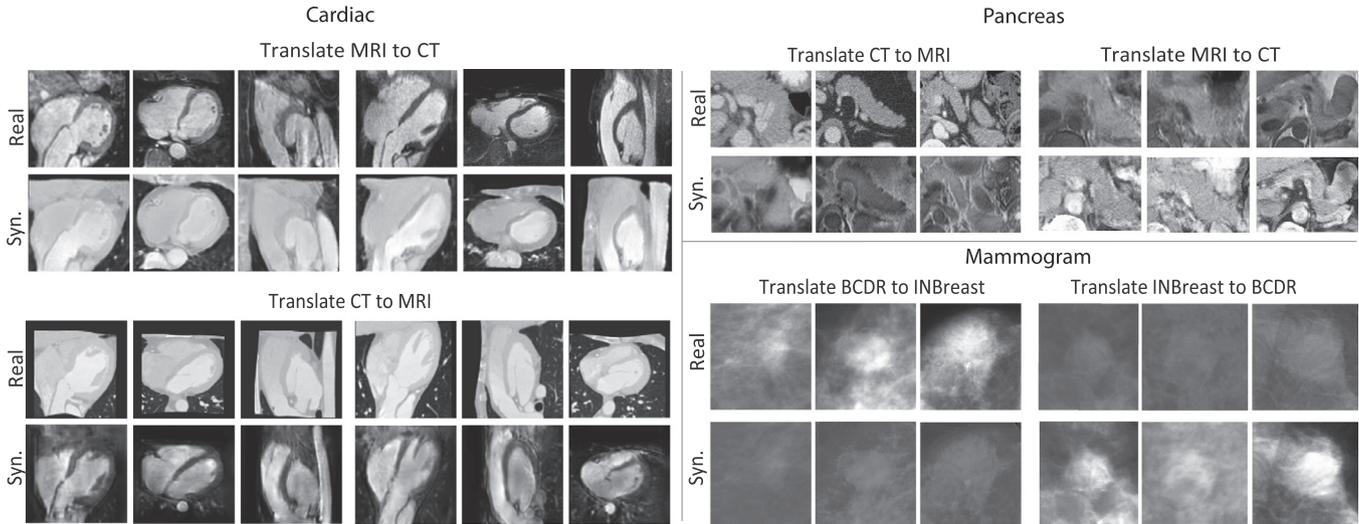
**Pancreatic datasets** Pancreas is one of the most challenging abdomen organs to segment for its shape variability and boundary vagueness. It is expensive to present machine learning methods with large size training data which is fully manual annotated. Aggregating data from multiple sources via modality translation may serve as a surrogate to this problem. To evaluate the proposed method in this circumstance, we prepared two datasets with 82 CT and 78 MRI scans, respectively. The CT dataset is collected from a publicly-available data source (Roth et al., 2016; Clark et al., 2013). In addition, the MRI dataset is collected from our in-house data source.

For pancreas datasets, we setup  $S_1$  by randomly selecting 40 cases from each of the CT and MRI datasets for model training and preserve the rests as testing data. We directly process them in 3D by cropping sub-volumes centered at the pancreas and have them resampled to the size of  $80 \times 64 \times 128$ .

**Mammogram datasets** Finally, we consider two publicly-available mammogram datasets, which are Breast Cancer Digital Repository (BCDR) (Lopez et al., 2012) and INbreast (Moreira et al., 2012; Dhungel et al., 2017). However, BCDR and INbreast are generated from different data sources and stored in different formats. According to our experience, the large appearance dependency makes two datasets domain isolated. Our method is useful to remove the barrier and make them benefit to each other. The BCDR contains 753 annotated lesion regions and the INbreast only has 116. Similar to the CT data in the cardiac and pancreatic dataset, we use images and annotations in BCDR to assist segmenting INbreast. To setup  $S_1$ , we randomly select 116 lesions from BCDR to match the number of lesions in INbreast. For both domains, 50% is used as training and validation and the rest 50% as testing data. We also set up  $S_2$ , which uses all the rest 637 BCDR lesions as the



**Fig. 4.** Illustration of the strategies to use synthetic data to improve segmentation. The left is the compared ad-hoc offline approach, which prepares the real and synthetic data separately and train a segmentation network. The right is our online approach: the real data in domain B is fed into the generator to synthesize domain A data. The domain A segmentor is trained with real and synthetic data jointly. The segmentor and generator cooperate with each other and minimize a joint objective.



**Fig. 5.** Qualitative results of the proposed method for image translation. On the left, we show three orthogonal cuts through the center of 3D cardiac volumes. On the top right, we show pancreas volumes of the maximum cross area. On the bottom right, we show real and synthetic lesion patches from mammography X-rays.

extra augmentation dataset. We crop subimages centered at the lesion regions with adequate padding, which is sized as 5% of image diagonal. We resize the subimages to  $1 \times 256 \times 256$  (i.e. a special 3D volume with unit depth).

## 5.2. Evaluation metrics

**Shape invariance evaluation** For methods of GANs to generate class-specific natural images, Salimans et al. (2016) proposes to use the Inception score to evaluate the diversity of generated images, by using an auxiliary trained classification network.

Inspired by this, we propose the S-core (segmentation score) to evaluate the shape invariance quality of synthetic images. We train two segmentation networks on the training data of respective modalities and compare the multi-class Dice score of synthetic volumes. For each synthetic volume, S-score is computed by comparing to the groundtruth of the corresponding real volume it is translated from. Hence, higher score indicates better-matched shape (i.e., less geometric distortion).

**Segmentation evaluation** Here we show how well our method can use the synthetic data and help improve segmentation. We compare to an ad-hoc approach as we mentioned above. Specifically, we individually train two segmentors, denoted as  $\tilde{S}_A$  and  $\tilde{S}_B$ . We treat the segmentation performance of them as Baseline (R) in the following. Then we train generators  $\tilde{G}_A$  and  $\tilde{G}_B$  with the adversarial and cycle-consistency losses (setting the weight of the shape-consistency loss to 0). Then by adding synthetic data, we perform the following comparison:

1. Ad-hoc approach (ADA): We use  $\tilde{G}_A$  and  $\tilde{G}_B$  to generate synthetic data (To make fair comparison, both synthetic data  $G_{A/B}(x_{B/A})$  and reconstructed data  $G_{A/B}(G_{B/A}(x_{A/B}))$  are used). We fine-tune  $\tilde{S}_{A/B}$  using synthetic together with real data (Fig. 4 left).<sup>4</sup>
2. Our method: We join  $\tilde{S}_A$ ,  $\tilde{S}_B$ ,  $\tilde{G}_A$ , and  $\tilde{G}_B$  (also with discriminators) and fine-tune the overall networks in an end-to-end fashion (Fig. 4 right), as specified in the training details.

## 5.3. Cross-domain translation evaluation

We evaluate the generators both qualitatively and quantitatively. Fig. 5 shows some typical synthetic results of our method for cardiac images, pancreatic images, and mammogram images. As can be observed visually, the synthetic images are close to real images and no obvious geometric distortion is introduced during image translation. Our method well preserves object anatomies, for example, aorta and spine in cardiac segmentation.

Table 2 shows the S-score of synthetic data from both domains for generators without the proposed shape-consistency loss, denoted as  $G$  w/o SC. We treat it as the baseline for comparison. Note that it is mostly similar with CycleGAN but using our optimized network designs. As can be seen, our method ( $G$  w/ SC) with shape-consistency achieves substantial improvement over the baseline for all the three datasets.

<sup>4</sup> At each training batch, we take half real and half synthetic data to prevent possible distraction from low-quality synthetic data.

**Table 2**

Shape quality evaluation using the proposed S-score (see text for definition) for synthesized images. The synthetic volumes using our method has much better shape quality on both modalities. SC denotes shape-consistency. In cardiac and pancreas datasets, A and B refer to CT and MRI, respectively. While in mammogram datasets, A and B refer to BCDR and INbreast, respectively.

Dataset	Method	S-score (%)	
		A	B
Cardiac	G w/o SC	66.8	67.5
	G w/ SC (Ours)	<b>69.2</b>	<b>69.6</b>
Pancreas	G w/o SC	58.5	53.4
	G w/ SC (Ours)	<b>62.3</b>	<b>59.4</b>
Mammo.	G w/o SC	69.8	64.5
	G w/ SC (Ours)	<b>72.6</b>	<b>68.9</b>

**Table 3**

The segmentation performance comparison. The baseline model is trained with only Real data (Baseline (R)), the second and third rows show the boosted results by using Synthetic data with the comparing ADA and our method, respectively.

Method	Dice score (%)	
	CT	MRI
Baseline (R)	67.8	70.3
ADA (R+S)	66.0	71.0
Ours (R+S)	<b>74.4</b>	<b>73.2</b>

5.4. Segmentation results

In the first experiment on  $S_1$ , we test the scenario that how well our method uses synthetic data to improve segmentation given only limited real data. Since we need to vary the number of data in one modality and fix another, we perform the experiments on both modalities, respectively.

Using the standard multi-class Dice metric (Dice, 1945), Table 3 compares the segmentation results for cardiac datasets. We use 14% real data and all synthetic data from the counter modality to train the U-Net segmentor and observe the proposed method achieves much better performance on both modalities. We notice ADA even deteriorates the performance. We speculate that it is because the baseline model trained with very few real data has not been stabilized. Synthetic data distracts optimization when used for training offline. While our method adapts them fairly well and leads to significant improvement.

Table 4 compares the segmentation results for pancreas. We use 25% real data and all synthetic data from the counter modality to train U-Net, PSP-Net, and Refine-Net and observe the proposed method achieves the best performance on both modalities for all segmentors. We find the worst cases in MRI have been largely relieved when synthetic data is introduced. For example, the worst case of Refine-Net is 0.01 Dice score meaning that the segmen-

**Table 4**

pancreas segmentation, we present the segmentation results with Dice score (%) in the form of  $\text{mean}_{-\min}^{+\max}$ .

Method	PSP-Net		U-Net		Refine-Net	
	CT	MRI	CT	MRI	CT	MRI
Baseline (R)	60.0 <sup>+74.6</sup> <sub>-21.1</sub>	47.7 <sup>+66.7</sup> <sub>-0.00</sub>	75.8 <sup>+88.3</sup> <sub>-18.4</sub>	64.1 <sup>+78.5</sup> <sub>-16.0</sub>	77.8 <sup>+87.8</sup> <sub>-11.4</sub>	66.1 <sup>+81.5</sup> <sub>-1.11</sub>
ADA (R+S)	61.3 <sup>+75.9</sup> <sub>-19.7</sub>	53.5 <sup>+67.1</sup> <sub>-12.1</sub>	76.8 <sup>+87.9</sup> <sub>-22.1</sub>	66.1 <sup>+80.3</sup> <sub>-28.6</sub>	78.2 <sup>+89.2</sup> <sub>-33.7</sub>	70.0 <sup>+81.2</sup> <sub>-27.6</sub>
Ours (R+S)	<b>61.5</b> <sup>+74.1</sup> <sub>-14.4</sub>	<b>53.8</b> <sup>+66.1</sup> <sub>-12.9</sub>	<b>77.2</b> <sup>+88.1</sup> <sub>-23.0</sub>	<b>67.3</b> <sup>+82.2</sup> <sub>-32.7</sub>	<b>78.8</b> <sup>+88.4</sup> <sub>-26.7</sub>	<b>70.4</b> <sup>+80.8</sup> <sub>-31.4</sub>

**Table 5**

Mammography segmentation.

	Method	Dice score (%)	
		BCDR	INbreast
PSP	Baseline (R)	54.3	66.6
	ADA (R+S)	<b>63.2</b>	69.6
	Ours (R+S)	62.8	<b>71.6</b>
Unet	Baseline (R)	62.2	66.2
	ADA (R+S)	68.2	70.6
	Ours (R+S)	<b>68.3</b>	<b>76.4</b>
Refine	Baseline (R)	72.6	76.5
	ADA (R+S)	74.9	80.4
	Ours (R+S)	<b>75.0</b>	<b>81.1</b>

tor fails to allocate pancreas. While, as augmented by the synthetic data, the worst case in our method has been improved to 0.31 Dice score presenting a much lower missed diagnosis rate.

It worth to notice that the reported score is lower than some pancreatic segmentation methods on these two datasets (Roth et al., 2018; Zhu et al., 2017b; Cai et al., 2018). It is because we use a smaller input volume size and a compact U-net configuration for GPU memory consideration. Our goal here is to show the relative improvement with our method. However, as we validated, a full-sized U-Net with our implementation trained with full resolution CT and MRI images achieves 85.3% and 79.9% Dice scores, respectively. The current comparable state-of-the-art for CT is 84.6% (Zhu et al., 2017b) and for MRI is 80.7% (Cai et al., 2018). Therefore, we argue that our method can straightforwardly achieve state-of-the-art performance on these two datasets.

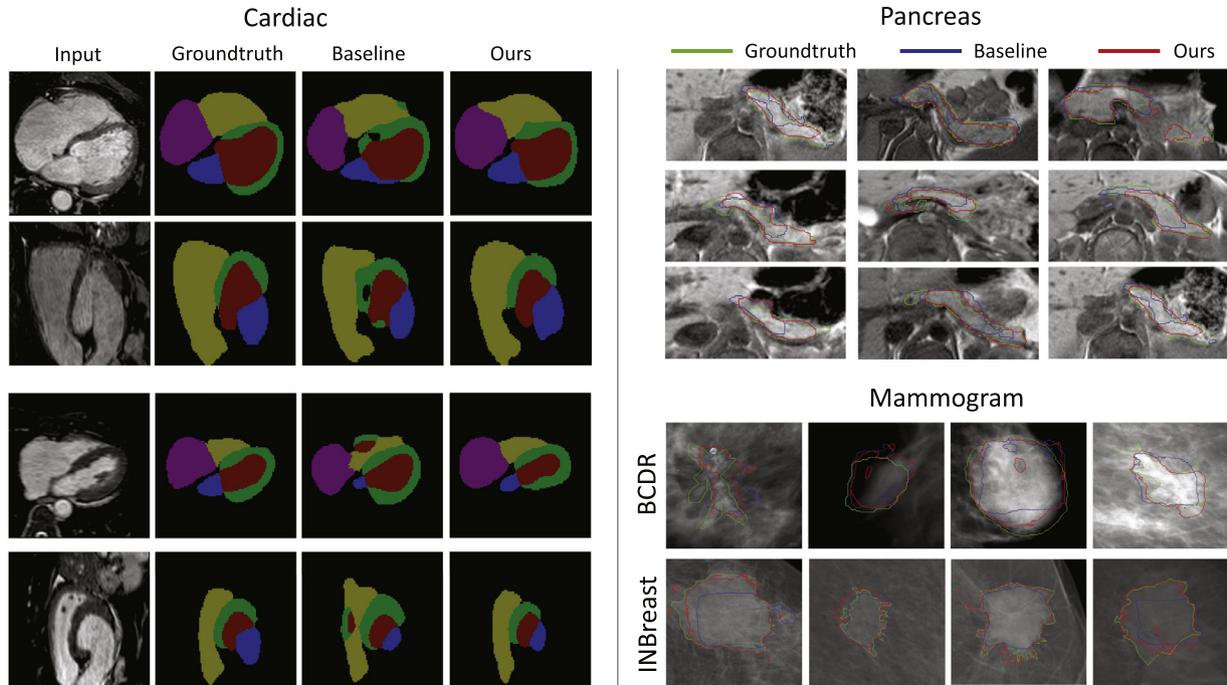
Table 5 presents the results for mammogram lesion segmentation. We use 14% of INbreast and 48% of BCDR (the result would be unstable if fewer BCDR images are used for model training). In the case of U-Net, we observe that our method significantly outperforms its baseline by 10% for INbreast segmentation. Although our result is not directly comparable with the state-of-the-art (Dhungel et al., 2017), we notice that synthetic training data systematically improves both BCDR and INBreast segmentation from baseline demonstrating the importance of including new training data for mammogram lesion segmentation and the effectiveness of the proposed image translation method.

The two datasets have large appearance dependency. A direct mixture of both datasets does not produce better segmentation accuracy. We observe 4% Dice score loss on INbreast segmentation when a direct combination of both datasets is used for segmentor training.

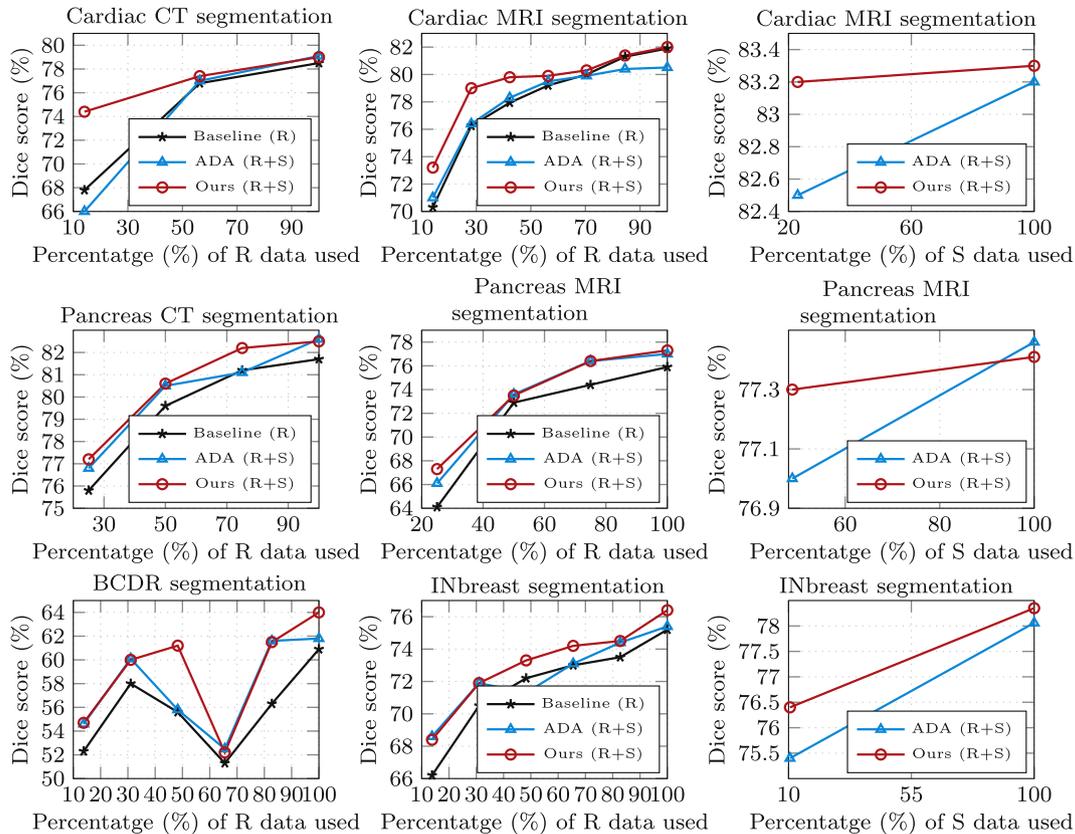
Finally, we demonstrate the qualitative segmentation results of our method in Fig. 6. By only using extra synthetic data, our method largely corrects the segmentation errors. We observe our method produces better-shaped results than its baseline. For example, as shown in the top right of Fig. 6, pancreas tail parts are well segmented in our results but lost in the baseline results.

5.5. Experimental analysis

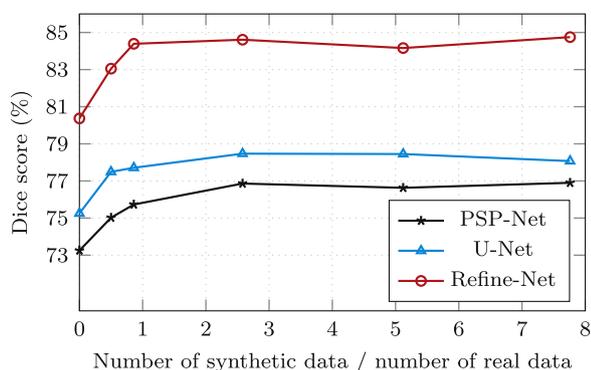
**Effectiveness of using synthetic data** We show the results by varying the number of real data used in Fig. 7 (left and mid-



**Fig. 6.** The qualitative evaluation of segmentation results on MRI. We show the axial and sagittal views of 2 cardiac samples, the axial view of 9 MRI pancrease samples and segmentations of 8 mammogram lesions on the left, top right and bottom right, respectively. Our method boosts the baseline segmentation network with only extra synthetic data. As can be observed, the segmentation errors of the baseline are largely corrected. Best viewed in color.



**Fig. 7.** The segmentation accuracy (mean Dice score) comparison to demonstrate the effectiveness of our method of using Synthetic data to boost segmentation. The left plot shows the segmentation accuracy by varying the percentage of Real data used for training segmentation on CT (BCDR) using dataset  $S_1$ , using an equal number of synthetic data. Baseline (R) is trained with only real data. Others are trained from it, e.g., ADA (R+S) is trained by adding only S data. The middle plot shows the same experiments on MRI (INbreast). The right plot shows results by varying the number of synthetic data on MRI (INbreast) using dataset  $S_2$  using a equal number of real data. Our method has consistently better performance. See text for details about comparing methods.



**Fig. 8.** The effectiveness of synthetic mammogram images with different segmentors. On the three segmentors, we repeat the experiment that each segmentor is first trained with BCDR images presenting the baseline Dice score. We then add synthetic data upto 8 times of the real data and observe the Dice scores.

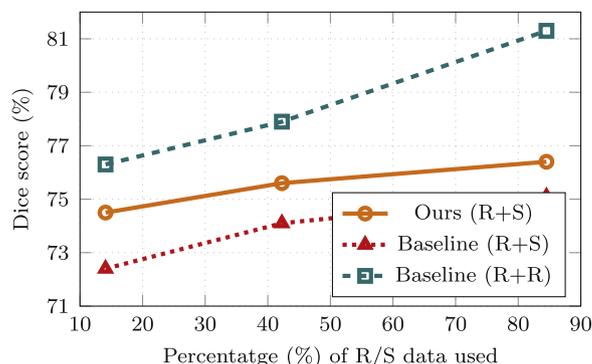
dle columns). Our method exhibits consistently better performance than the ADA. In addition, we notice the increment is growing slower as the number of real data increases. We hypothesize that is that more real data makes the segmentors get closer to its capacity, so the effect of extra synthetic data gets smaller. But this situation can be definitely balanced out by increasing the size of segmentors with sufficient GPU memory.

We then apply the experiment on  $S_2$ , which has much more CT (BCDR) data, so we aim at boosting the MRI (INbreast) segmentor. We vary the number of used synthetic data and use all real MRI (INbreast) data. Fig. 7 (right column) compares the results. Our method still shows better performance. As can be observed, for cardiac segmentation, our method uses 23% synthetic data to reach the accuracy of the ADA when it uses 100% synthetic data. For mammogram lesion segmentation, our method outperforms ADA in all cases.

With the  $S_2$  mammogram dataset, we present the second experiment on investigating synthetic data with different segmentors including U-Net, PSP-Net, and Refine-Net. Here, we first train each segmentor on 100% real INbreast training images to generate baseline Dice scores. With these pre-trained segmentors, we then fine-tune them with the proposed method on extended training sets, which contains synthetic images up to 8 times of the number of real images. In Fig. 8, we observe our method outperforms the corresponding baselines by large margins. For example, our method improves Refined-Net from 80.4 to 84.8 (Dice scores). For all of the three segmentors, we notice the major improvement is gained by adding the amount of synthetic data to as many as 3 times of the real data. To explain this, we argue that the segmentors may have reached the upper bound of their capacities. To represent the effectiveness of synthetic data of larger data size, it may require stronger segmentors.

**Gap between synthetic and real data** Reducing the distribution gap between real and synthetic data is the key to make synthetic data useful for segmentation. Here we show a way to interpret the gap between synthetic and real data by evaluating their performance to improve segmentation in terms of cardiac datasets. On dataset  $S_1$ , we train a MRI segmentor using 14% real data. Then we boost the segmentor by adding 1) pure MRI real data, 2) using ADA, and 3) using our method. As shown in Fig. 9, our method reduces the gap of the ADA significantly, *i.e.*, by 61% given 14% real data and 20.9% given 85% real data.

Moreover, we find that when using the synthetic data as augmented data offline (our comparing baseline), too much synthetic data could diverge the network training. While in our method, we did not observe such situation. However, we also observe that the gap is more difficult to reduce the number of reading data in-



**Fig. 9.** The gap analysis of Real and Synthetic cardiac data. For all comparing methods, we use one pre-trained network with 14% real data, whose Dice score is 70.3%. Then we vary the number of R or S data used to boost segmentation of Baseline (R+R), Baseline (R+S), and Ours (R+S). Our method significantly reduces the gap for all settings.

creases. Although one of the reasons is due to the modal capacity, we believe the solution of this *gap-reduction* worth further study.

## 6. Conclusion

In this paper, we present a novel method towards cross-modal medical 3D/2D images translation and segmentation, which are two significant tasks in medical imaging. We address three key problems that are important in synthesizing realistic medical images: 1) learn from unpaired data, 2) keep anatomy (*i.e.* shape) consistency, and 3) use synthetic data to improve volume segmentation effectively. Acquiring large annotated data for each of medical imaging domains is very expensive. Our method is valuable to reduce the isolation in different modalities and make them beneficial to each other for domain adaptation and segmentation. With extensive experiments on datasets of three different medical imaging modalities, we validate the effectiveness and superiority of the proposed method.

## References

- Arjovsky, M., Chintala, S., Bottou, L., 2017. Wasserstein gan. arXiv:1701.07875.
- Berthelot, D., Schumm, T., Metz, L., 2017. Began: Boundary equilibrium generative adversarial networks. arXiv:1703.10717.
- Bousmalis, K., Silberman, N., Dohan, D., Erhan, D., Krishnan, D., 2017. Unsupervised pixel-level domain adaptation with generative adversarial networks. CVPR.
- Burgos, N., Cardoso, M.J., Guerreiro, F., Veiga, C., Modat, M., McClelland, J., Knopf, A.-C., Punwani, S., Atkinson, D., Arridge, S.R., et al., 2015. Robust ct synthesis for radiotherapy planning: application to the head and neck region. MICCAI.
- Cai, J., Lu, L., Xing, F., Yang, L., 2018. Pancreas segmentation in CT and MRI images via domain specific network designing and recurrent neural contextual learning. arXiv:1803.11303.
- Cao, X., Yang, J., Gao, Y., Guo, Y., Wu, G., Shen, D., 2017. Dual-core steered non-rigid registration for multi-modal images via bi-directional image synthesis. Med. Image Anal. 41, 18–31.
- Chen, M., Carass, A., Jog, A., Lee, J., Roy, S., Prince, J.L., 2017. Cross contrast multi-channel image registration using image synthesis for mr brain images. Med. Image Anal. 36, 2–14.
- Clark, K.W., Vendt, B.A., Smith, K., Freymann, J.B., Kirby, J.S., Koppel, P., Moore, S.M., Phillips, S.R., Maffitt, D.R., Pringle, M., Tarbox, L., Prior, F.W., 2013. The cancer imaging archive (TCIA): maintaining and operating a public information repository. J. Digital Imaging 26 (6), 1045–1057.
- Cordier, N., Delingette, H., L  , M., Ayache, N., 2016. Extended modality propagation: image synthesis of pathological cases. IEEE Trans. Med. Imaging 35 (12), 2598–2608.
- Costa, P., Galdran, A., Meyer, M.I., Niemeijer, M., Abr  moff, M., Mendon  sa, A.M., Campilho, A., 2018. End-to-end adversarial retinal image synthesis. IEEE Trans. Med. Imaging 37 (3), 781–791. doi:10.1109/TMI.2017.2759102.
- Dhungel, N., Carneiro, G., Bradley, A.P., 2017. A deep learning approach for the analysis of masses in mammograms with minimal user intervention. Med. Image Anal. 37, 114–128.
- Dice, L.R., 1945. Measures of the amount of ecologic association between species. Ecology 26 (3), 297–302.

- Gong, Y., Karanam, S., Wu, Z., Peng, K.-C., Ernst, J., Doerschuk, P. C., 2017. Learning compositional visual concepts with mutual consistency. arXiv:1711.06148.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets. NIPS.
- He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. CVPR.
- Huang, Y., Shao, L., Frangi, A.F., 2018. Cross-modality image synthesis via weakly coupled and geometry co-regularized joint dictionary learning. IEEE Trans. Med. Imaging 37 (3), 815–827.
- Huo, Y., Xu, Z., Bao, S., Assad, A., Abramson, R.G., Landman, B.A., 2018. Adversarial synthesis learning enables segmentation without target modality ground truth. ISBI.
- Iglesias, J.E., Konukoglu, E., Zikic, D., Glocker, B., Van Leemput, K., Fischl, B., 2013. Is synthesizing mri contrast useful for inter-modality analysis? MICCAI.
- Ioffe, S., Szegedy, C., 2015. Batch normalization: accelerating deep network training by reducing internal covariate shift. ICML.
- Isola, P., Zhu, J.-Y., Zhou, T., Efros, A.A., 2017. Image-to-image translation with conditional adversarial networks. CVPR.
- Kamnitsas, K., Baumgartner, C., Ledig, C., Newcombe, V., Simpson, J., Kane, A., Menon, D., Nori, A., Criminisi, A., Rueckert, D., et al., 2017. Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. IPML.
- Kim, T., Cha, M., Kim, H., Lee, J., Kim, J., 2017. Learning to discover cross-domain relations with generative adversarial networks. ICML.
- Kingma, D., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv:1412.6980.
- Kohl, S., Bonekamp, D., Schlemmer, H.-P., Yaqubi, K., Hohenfellner, M., Hadaschik, B., Radtke, J.-P., Maier-Hein, K., 2017. Adversarial networks for the detection of aggressive prostate cancer. arXiv:1702.08014.
- Lin, G., Milan, A., Shen, C., Reid, I.D., 2017. Refinenet: multi-path refinement networks for high-resolution semantic segmentation. In: CVPR, pp. 5168–5177.
- Liu, M.-Y., Breuel, T., Kautz, J., 2017. Unsupervised image-to-image translation networks. NIPS.
- Liu, M.-Y., Tuzel, O., 2016. Coupled generative adversarial networks. NIPS.
- Lopez, M.G., Moura, D.C., Pollán, R.R., Valiente, J.M.F., Ortega, C.S., Ramos, I., et al., 2012. BCDR: a breast cancer digital repository. In: In 15th International Conference on Experimental Mechanics.
- Luc, P., Couprie, C., Chintala, S., Verbeek, J., 2016. Semantic segmentation using adversarial networks. NIPS Workshop on Adversarial Training.
- Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Smolley, S.P., 2017. Least squares generative adversarial networks. In: Computer Vision (ICCV), 2017 IEEE International Conference on. IEEE, pp. 2813–2821.
- Moreira, I.C., Amaral, I., Domingues, I., Cardoso, A., Cardoso, M.J., Cardoso, J.S., 2012. Inbreast: toward a full-field digital mammographic database. Acad. Radiol. 19 (2), 236–248.
- Nie, D., Trullo, R., Lian, J., Wang, L., Petitjean, C., Ruan, S., Wang, Q., Shen, D., 2018. Medical image synthesis with deep convolutional adversarial networks. IEEE Trans. Biomed. Eng. doi:10.1109/TBME.2018.2814538.
- Odena, A., Dumoulin, V., Olah, C., 2016. Deconvolution and checkerboard artifacts. Distill.
- Osokin, A., Chessel, A., Salas, R.E.C., Vaggi, F., 2017. Gans for biological image synthesis. ICCV.
- Peng, X., Tang, Z., Yang, F., Feris, R.S., Metaxas, D., 2018. Jointly optimize data augmentation and network training: adversarial data augmentation in human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2226–2234.
- Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: convolutional networks for biomedical image segmentation. MICCAI.
- Roth, H.R., Farag, A., Turkbey, E.B., Lu, L., Liu, J., Summers, R.M., 2016. Data from pancreas-ct. the cancer imaging archive.
- Roth, H.R., Lu, L., Lay, N., Harrison, A.P., Farag, A., Sohn, A., Summers, R.M., 2018. Spatial aggregation of holistically-nested convolutional neural networks for automated pancreas localization and segmentation. Med. Image Anal. 45, 94–107.
- Roth, H.R., Lu, L., Seff, A., Cherry, K.M., Hoffman, J., Wang, S., Liu, J., Turkbey, E., Summers, R.M., 2014. A new 2.5 d representation for lymph node detection using random sets of deep convolutional neural network observations. In: MICCAI, pp. 520–527.
- Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X., 2016. Improved techniques for training gans. NIPS.
- Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., Webb, R., 2017. Learning from simulated and unsupervised images through adversarial training. CVPR.
- Ulyanov, D., Vedaldi, A., Lempitsky, V., 2016. Instance normalization: the missing ingredient for fast stylization. arXiv:1607.08022.
- Van Nguyen, H., Zhou, K., Vemulapalli, R., 2015. Cross-domain synthesis of medical images using efficient location-sensitive deep network. MICCAI.
- Vemulapalli, R., Van Nguyen, H., Kevin Zhou, S., 2015. Unsupervised cross-modal synthesis of subject-specific scans. ICCV.
- Xue, Y., Xu, T., Zhang, H., Long, R., Huang, X., 2017. Segan: adversarial network with multi-scale  $L_1$  loss for medical image segmentation. arXiv:1706.01805.
- Yang, D., Xu, D., Zhou, S.K., Georgescu, B., Chen, M., Grbic, S., Metaxas, D., Comaniciu, D., 2017. Automatic liver segmentation using an adversarial image-to-image network. MICCAI.
- Yi, Z., Zhang, H., Gong, P.T., et al., 2017. Dualgan: unsupervised dual learning for image-to-image translation. ICCV.
- Zbontar, J., Knoll, F., Sriram, A., Muckley, M. J., Bruno, M., Defazio, A., Parente, M., Geras, K. J., Katsnelson, J., Chandarana, H., et al., 2018. fastmri: An open dataset and benchmarks for accelerated MRI. arXiv:1811.08839.
- Zhang, R., Isola, P., Efros, A.A., 2016. Colorful image colorization. ECCV.
- Zhang, Z., Chen, P., Sapkota, M., Yang, L., 2017a. Tandemnet: distilling knowledge from medical images using diagnostic reports as optional semantic references. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 320–328.
- Zhang, Z., Xie, Y., Xing, F., McGough, M., Yang, L., 2017b. Mdnet: a semantically and visually interpretable medical image diagnosis network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6428–6436.
- Zhang, Z., Xie, Y., Yang, L., 2018a. Photographic text-to-image synthesis with a hierarchically-nested adversarial network. CVPR.
- Zhang, Z., Yang, L., Zheng, Y., 2018b. Translating and segmenting multimodal medical volumes with cycle- and shape-consistency generative adversarial network. CVPR.
- Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid scene parsing network. In: CVPR, pp. 6230–6239.
- Zhao, L., Peng, X., Kapadia, M., Metaxas, D.N., 2018. Learning to forecast and refine residual motion for image-to-video generation. In: European Conference on Computer Vision.
- Zhu, J.-Y., Park, T., Isola, P., Efros, A.A., 2017a. Unpaired image-to-image translation using cycle-consistent adversarial networks. ICCV.
- Zhu, Z., Xia, Y., Shen, W., Fishman, E.K., Yuille, A.L., 2017b. A 3d coarse-to-fine framework for automatic pancreas segmentation. CoRR abs/1712.00201. 1712.00201.