Towards Cross-modal Organ Translation and Segmentation: A Cycle- and Shape-Consistent Generative Adversarial Network[1]

Jinzheng Cai[a,2], Zizhao Zhang[a], Lei Cui[b], Yefeng Zheng[c], Lin Yang[a]

[a] University of Florida
[b] Northwest University
[c] Siemens Healthcare

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.

1 Introduction

In current clinical practice, multiple imaging modalities may be available for disease diagnosis and surgical planning [6, 7]. 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 [4].

Neural networks have been widely used for medical image analysis, such as detection [39], segmentation [36, 38], and classification [50, 51]. 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 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.

[1] Preprint, to appear in Medical Image Analysis (https://doi.org/10.1016/j.media.2018.12.002)
[2] E-mail: caijinzhengcn@gmail.com

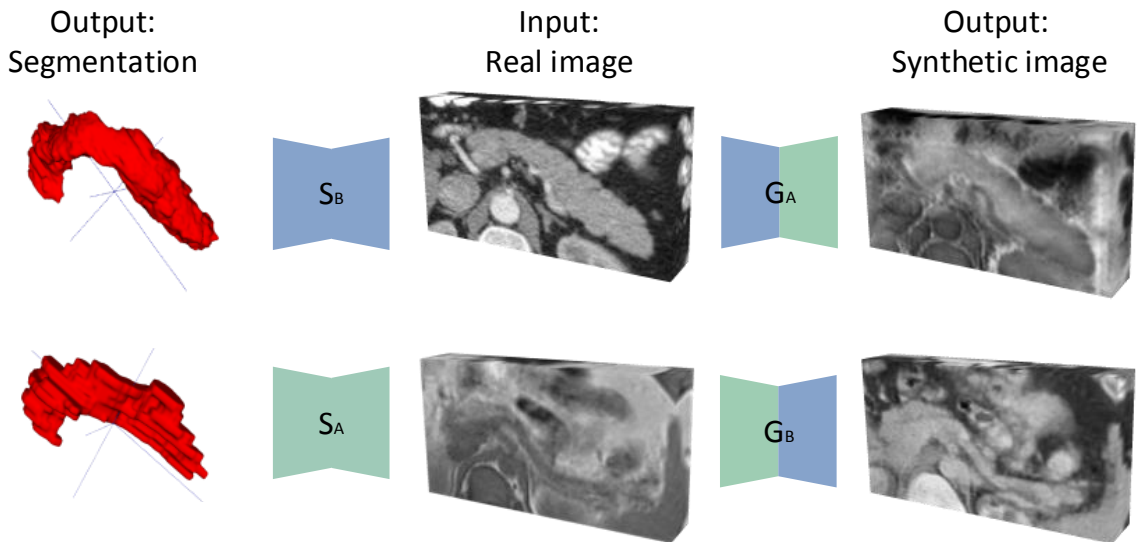

Figure 1. Our method learns two parallel sets of generators $G_{A/B}$ and segmentors $S_{A/B}$ for two modalities A and B to translate and segment holistic 3D volumes. Here we illustrate using CT and MRI pancreatic 3D images. Best viewed in color.

To synthesize medical images, recent advances [10, 32] have used generative adversarial networks (GANs) [14] 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 [21, 48], 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 [3, 41].

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 [56] 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 [53], 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 [54], U-Net [36], and Refine-Net [25] 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 [16, 32], and the second is to use synthetic data from other modalities with sufficient labeled data to help classification tasks (*e.g.*, domain adaption [21]).

In computer vision, recent image-to-image translation is formulated as a pixel-to-pixel mapping using encoder-decoder CNNs [13, 20, 22, 26, 56]. Several studies have explored cross-modality translation for medical images, using probabilistic generative model [9], sparse coding [16, 44], GANs [32, 34], CNN [43], etc. GANs have attracted wide interests in helping addressing such tasks to generate high-quality, less blurry results [1, 2, 14, 52]. More recent studies apply pixel-to-pixel GANs for brain MRI to CT image translation [21, 32] and retinal vessel annotation to image translation [10]. 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 [27, 44].

Synthesizing medical data to overcome insufficient labeled data attracted wide interests recently [17, 18, 35, 41, 55]. Due to the diversity of medical modalities, learning an unsupervised translation between modalities is a promising direction [10]. [21] 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 [24, 29, 45, 46] 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. Convolutional neural network (CNN) in nature needs to be feed 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

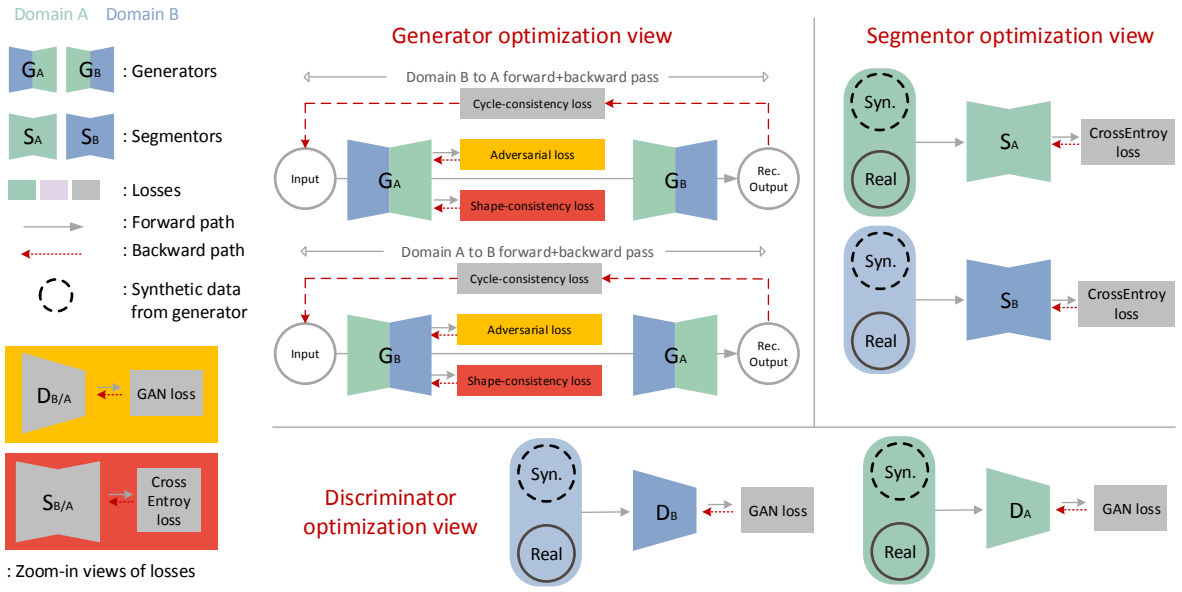

Figure 2. The illustration of our method from the generator optimization view (top-left), the segmentor optimization view (top-right), and the discriminator optimization view (bottom). The left panel illustrates each architecture components. Domain A and Domain B are illustrated using two colors. **Generator 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.

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 [49]. ConditionalGAN [20] 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 [56] and other similar methods [22, 47] 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 [14].

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,

$$G_A^{'} = G_A \circ T \text{ and } G_B^{'} = G_B \circ T^{-1} \tag{1}$$

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 being 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:

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

where x_A is a sample from domain A and x_B is from domain B. $p_d()$ denotes the data distribution at a certain domain. $\mathcal{L}_{cyc}$ uses the L1 loss over all voxels (or pxiels), 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.*, $S_A : A \to Y$ and $S_B : B \to Y$, 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

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

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)

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

and define our shape-consistency loss as

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

where $y_A^i, y_B^i \in \{0, 1, ..., C\}$ 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. [41] 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:

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

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

$$\begin{aligned}\mathcal{L}_{GAN}(G_B, D_B) = \mathbb{E}_{x_B\sim p_{\text{d}}(x_B)}\Big[||D_B(x_B)||_2\Big] \\ + \mathbb{E}_{x_A\sim p_{\text{d}}(x_A)}\Big[||D_B(G_B(x_A)) - \mathbf{1}||_2\Big]\end{aligned} \tag{7}$$

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 [20, 56]. We follow the Least Squires GAN (LSGAN) [30] loss to replace the original cross-entropy loss in GANs with the simple mean squared loss. λ is set to 10 and γ 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.

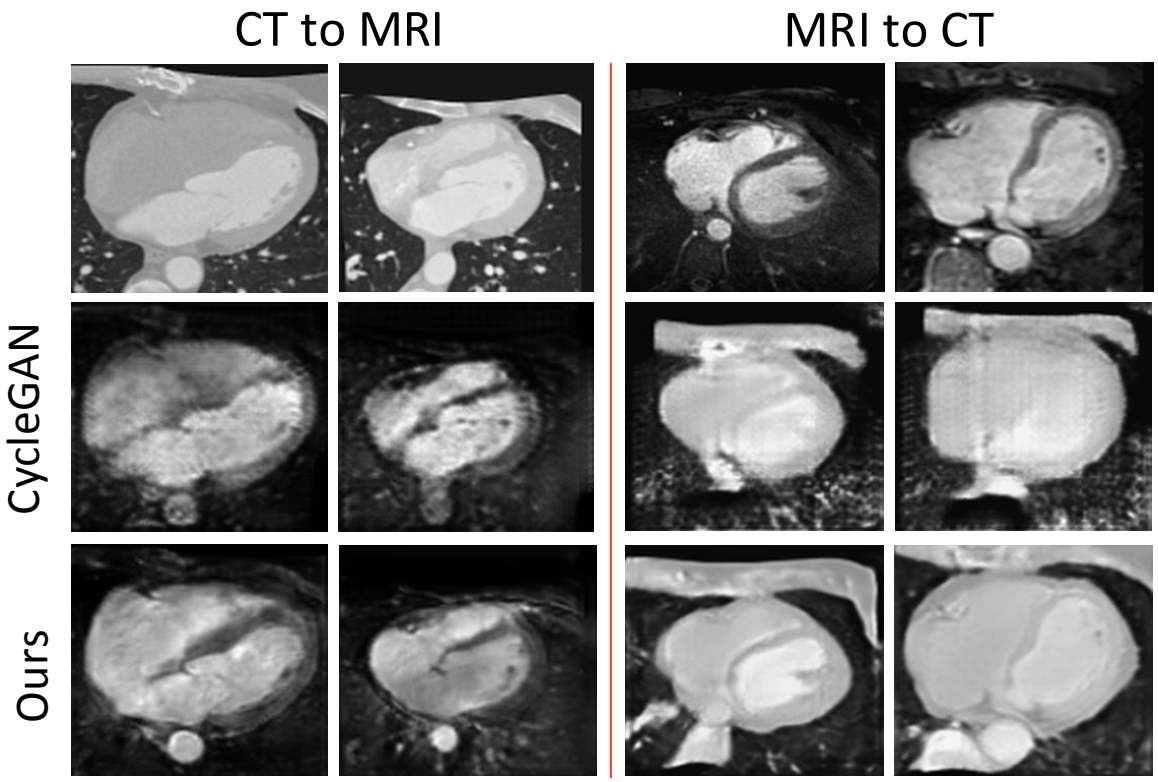

Figure 3. Example outputs on 2D slides of 3D cardiovascular CT and MRI images of the results using 3D CycleGAN (second row) and ours (third row). The first row is the input samples. The original results of CycleGAN have severe artifacts, checkerboard effects, and missing anatomies (*e.g.*, descending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

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 [39] or sub-volumes [21], 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 [42] (performs better than batch normalization [19]) and ReLU for generators or LeakyReLU for discriminators. CycleGAN originally designs generators with multiple residual blocks [15]. 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 [36] 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 3×3×3 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 3×3×3 convolution to

realize upsampling as well as channel changes. It is also observed in [33] 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[3].

For discriminators, we adopt the PatchGAN proposed by [41] to classify whether an overlapping sub-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 [36], Refine-Net [25], and PSP-Net [54] 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-poling 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 $3\times3\times3$ 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 [23] 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. 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

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

G_A	D_A
Inputs(b×80×128×128×1)	Inputs(b×80×128×128×1)
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))	

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 concatention 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).

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 4,354 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 86×112×112 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 $\mathcal{S}_1$ and $\mathcal{S}_2$. For $\mathcal{S}_1$, 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 $\mathcal{S}_2$, we use all the rest 4,212 CT volumes as an extra augmentation dataset, which is used to generate synthetic MRI volumes for segmentation. We fix the testing data in $\mathcal{S}_1$ 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

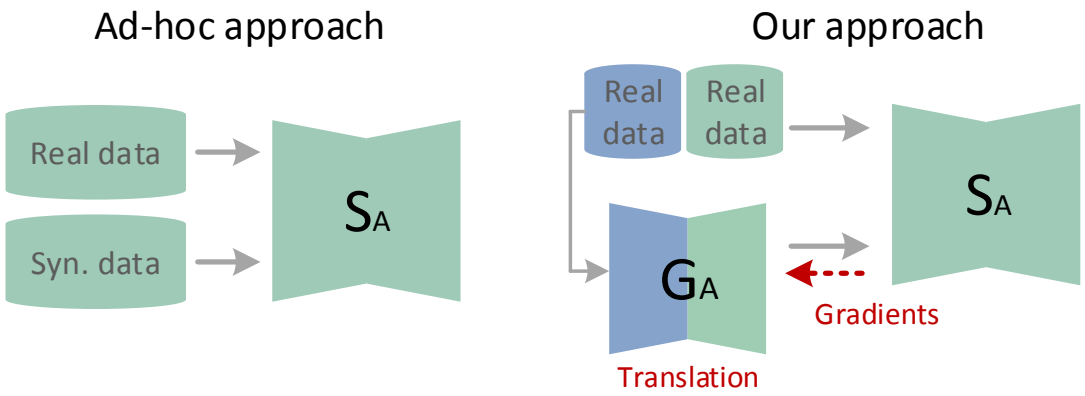

Figure 4. Illustration of the strategies to use synthetic data to improve segmentation. The left is the compared ad-hoc offline approach, which prepares the real and synthetic data separately and train a segmentation network. The right is our online approach: the real data in domain B is fed into the 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.

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-availabel data source [8,37]. In addition, the MRI dataset is collected from our in-house data source.

For pancreas datasets, we setup $\mathcal{S}_1$ 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 $80\times64\times128$.

Mammogram Datasets Finally, we consider two publicly-available mammogram datasets, which are Breast Cancer Digital Repository (BCDR) [28] and INbreast [11,31]. 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 $\mathcal{S}_1$, 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 $\mathcal{S}_2$, which uses all the rest 637 BCDR lesions as the 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, [40] 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

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

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.

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)[4].

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.

5.4 Segmentation Results

In the first experiment on $\mathcal{S}_1$, 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 [12], 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

[4] At each training batch, we take half real and half synthetic data to prevent possible distraction from low-quality synthetic data.

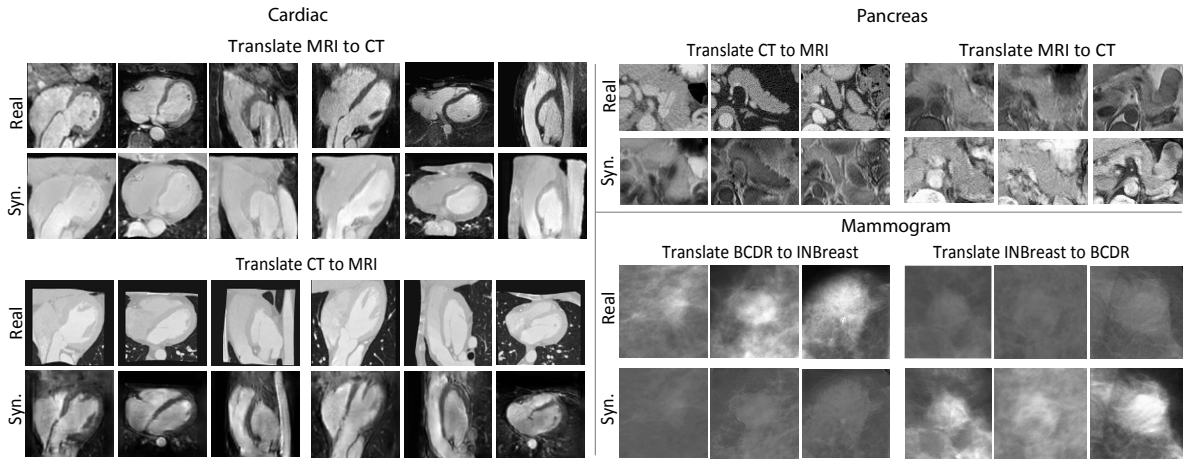

Figure 5. Qualitative results of the proposed method for image translation. On the left, we show three orthogonal cuts through the center of 3D cardiac volumes. On the top right, we show pancreas volumes of the maximum cross area. On the bottom right, we show real and synthetic lesion patches from mammography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

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

Table 3. The segmentation performance comparison. The baseline model is trained with only **R**eal data (Baseline (R)), the second and third rows show the boosted results by using **S**ynthetic data with the comparing ADA and our method, respectively.

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

Table 4. Pancreas segmentation, we present the segmentation results with Dice score (%) in the form of $\text{mean}^{+\text{max}}_{-\text{min}}$.

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 segmentor 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 [5,38,57]. It is because we use a smaller input volume size and a compact U-net configuration

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

Table 5. Mammography segmentation.

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% [57] and for MRI is 80.7% [5]. 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 [11], 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 middle 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 $\mathcal{S}_2$, 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 $\mathcal{S}_2$ 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,

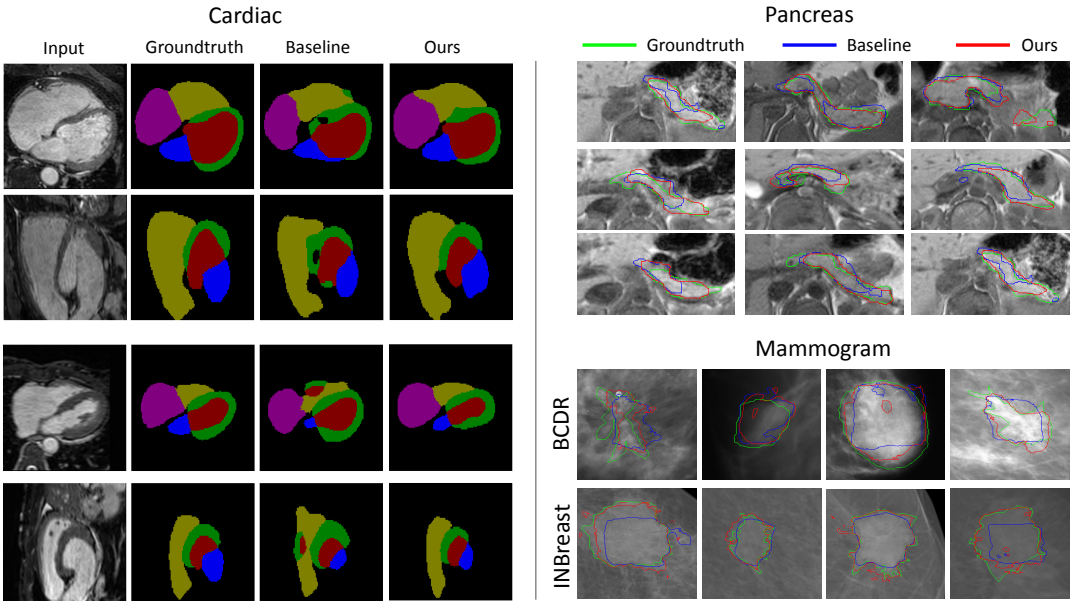

Figure 6. The qualitative evaluation of segmentation results on MRI. We show the axial and sagittal views of 2 cardiac samples, the axial view of 9 MRI pancrease samples and segmentations of 8 mammogram lesions on the left, top right and bottom right, respectively. Our method boosts the baseline segmentation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

we then finetune 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 $\mathcal{S}_1$, we train an 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 increases. 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

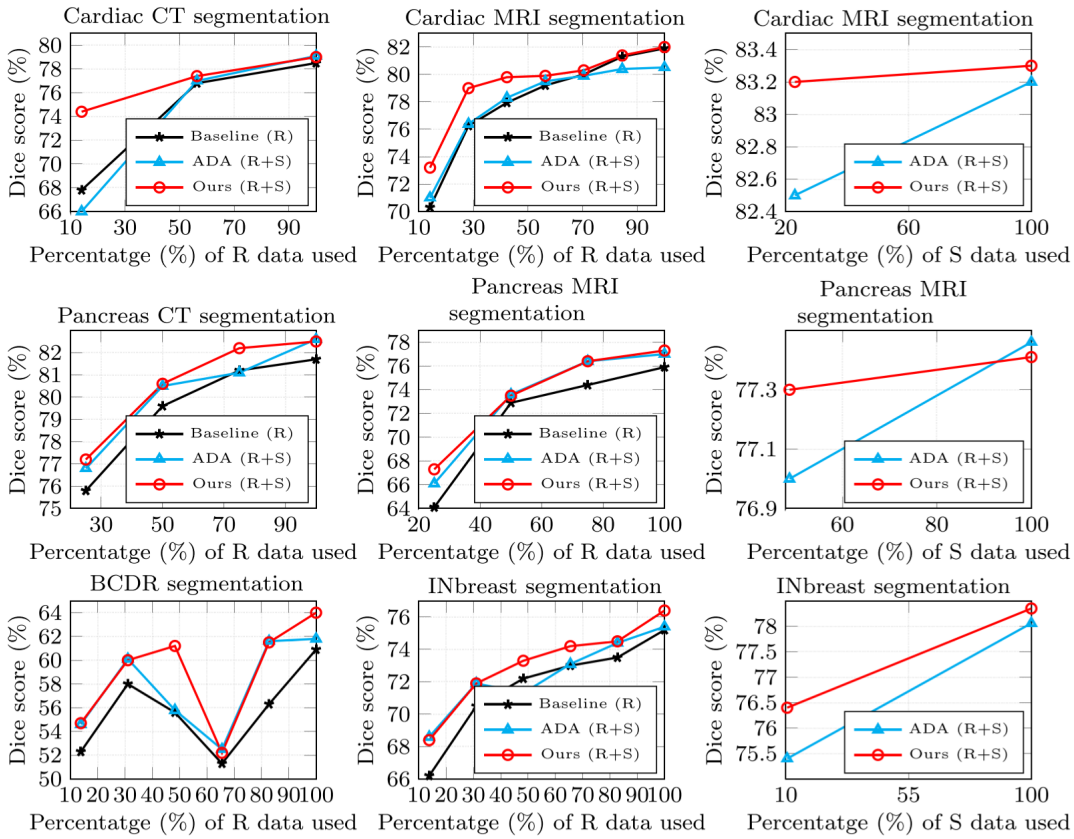

Figure 7. The segmentation accuracy (mean Dice score) comparison to demonstrate the effectiveness of our method of using **S**ynthetic data to boost segmentation. The left plot shows the segmentation accuracy by varying the percentage of **R**eal data used for training segmentation on CT (BCDR) using dataset $\mathbf{S}_1$, using a 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 $\mathbf{S}_2$ using a equal number of real data. Our method has consistently better performance. See text for details about comparing methods.

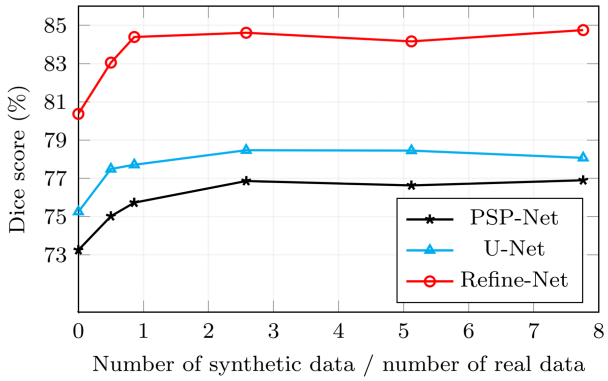

Figure 8. The effectiveness of synthetic mammogram images with different segmentors. On the three segmentors, we repeat the experiment that each segmentor is first trained with BCDR images presenting the baseline Dice score. We then add synthetic data upto 8 times of the real data and observe the Dice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

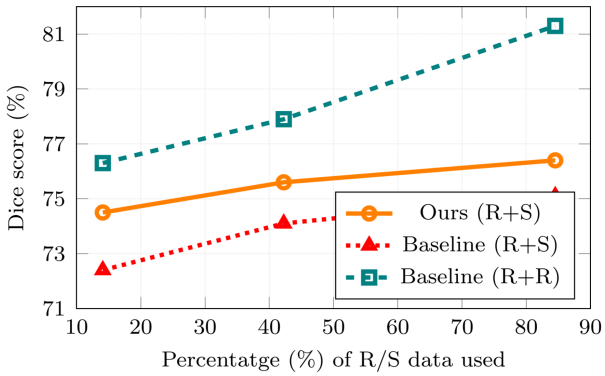

Figure 9. The gap analysis of **R**eal and **S**ynthetic cardiac data. For all comparing methods, we use one pre-trained network with 14% real data, whose Dice score is 70.3%. Then we vary the number of R or S data used to boost segmentation of Baseline (R+R), Baseline (R+S), and Ours (R+S). Our method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

(*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

1. M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein gan. arXiv preprint arXiv:1701.07875, 2017.
2. D. Berthelot, T. Schumm, and L. Metz. Began: Boundary equilibrium generative adversarial networks. arXiv preprint arXiv:1703.10717, 2017.
3. K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. CVPR, 2017.
4. N. Burgos, M. J. Cardoso, F. Guerreiro, C. Veiga, M. Modat, J. McClelland, A.-C. Knopf, S. Punwani, D. Atkinson, S. R. Arridge, et al. Robust ct synthesis for radiotherapy planning: application to the head and neck region. In MICCAI, 2015.
5. J. Cai, L. Lu, F. Xing, and L. Yang. Pancreas segmentation in CT and MRI images via domain specific network designing and recurrent neural contextual learning. arXiv preprint arXiv:1803.11303, 2018.
6. X. Cao, J. Yang, Y. Gao, Y. Guo, G. Wu, and D. Shen. Dual-core steered non-rigid registration for multi-modal images via bi-directional image synthesis. Medical Image Analysis, 41:18 – 31, 2017.
7. M. Chen, A. Carass, A. Jog, J. Lee, S. Roy, and J. L. Prince. Cross contrast multi-channel image registration using image synthesis for mr brain images. Medical Image Analysis, 36:2 – 14, 2017.
8. K. W. Clark, B. A. Vendt, K. Smith, J. B. Freymann, J. S. Kirby, P. Koppel, S. M. Moore, S. R. Phillips, D. R. Maffitt, M. Pringle, L. Tarbox, and F. W. Prior. The cancer imaging archive (TCIA): maintaining and operating a public information repository. J. Digital Imaging, 26(6):1045–1057, 2013.
9. N. Cordier, H. Delingette, M. Lê, and N. Ayache. Extended modality propagation: Image synthesis of pathological cases. IEEE Transactions on Medical Imaging, 35(12):2598–2608, Dec 2016.

10. P. Costa, A. Galdran, M. I. Meyer, M. Niemeijer, M. Abràmoff, A. M. Mendonça, and A. Campilho. End-to-end adversarial retinal image synthesis. IEEE Transactions on Medical Imaging, 37(3):781–791, March 2018.

11. N. Dhungel, G. Carneiro, and A. P. Bradley. A deep learning approach for the analysis of masses in mammograms with minimal user intervention. Medical Image Analysis, 37:114 – 128, 2017.

12. L. R. Dice. Measures of the amount of ecologic association between species. Ecology, 26(3):297–302, 1945.

13. Y. Gong, S. Karanam, Z. Wu, K.-C. Peng, J. Ernst, and P. C. Doerschuk. Learning compositional visual concepts with mutual consistency. arXiv preprint arXiv:1711.06148, 2017.

14. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In NIPS, 2014.

15. K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In CVPR, 2016.

16. Y. Huang, L. Shao, and A. F. Frangi. Cross-modality image synthesis via weakly coupled and geometry co-regularized joint dictionary learning. IEEE Transactions on Medical Imaging, 37(3):815–827, March 2018.

17. Y. Huo, Z. Xu, S. Bao, A. Assad, R. G. Abramson, and B. A. Landman. Adversarial synthesis learning enables segmentation without target modality ground truth. ISBI, 2018.

18. J. E. Iglesias, E. Konukoglu, D. Zikic, B. Glocker, K. Van Leemput, and B. Fischl. Is synthesizing mri contrast useful for inter-modality analysis? In MICCAI, 2013.

19. S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In ICML, 2015.

20. P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. CVPR, 2017.

21. K. Kamnitsas, C. Baumgartner, C. Ledig, V. Newcombe, J. Simpson, A. Kane, D. Menon, A. Nori, A. Criminisi, D. Rueckert, et al. Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. In IPMI, 2017.

22. T. Kim, M. Cha, H. Kim, J. Lee, and J. Kim. Learning to discover cross-domain relations with generative adversarial networks. In ICML, 2017.

23. D. Kingma and J. Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

24. S. Kohl, D. Bonekamp, H.-P. Schlemmer, K. Yaqubi, M. Hohenfellner, B. Hadaschik, J.-P. Radtke, and K. Maier-Hein. Adversarial networks for the detection of aggressive prostate cancer. arXiv preprint arXiv:1702.08014, 2017.

25. G. Lin, A. Milan, C. Shen, and I. D. Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In CVPR, pages 5168–5177, 2017.

26. M.-Y. Liu, T. Breuel, and J. Kautz. Unsupervised image-to-image translation networks. In NIPS, 2017.

27. M.-Y. Liu and O. Tuzel. Coupled generative adversarial networks. In NIPS, 2016.

28. M. G. Lopez, D. C. Moura, R. R. Pollán, J. M. F. Valiente, C. S. Ortega, I. Ramos, et al. BCDR: a breast cancer digital repository. In In 15th International Conference on Experimental Mechanics., 2012.

29. P. Luc, C. Couprie, S. Chintala, and J. Verbeek. Semantic segmentation using adversarial networks. NIPS Workshop on Adversarial Training, 2016.

30. X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. P. Smolley. Least squares generative adversarial networks. In Computer Vision (ICCV), 2017 IEEE International Conference on, pages 2813–2821. IEEE, 2017.

31. I. C. Moreira, I. Amaral, I. Domingues, A. Cardoso, M. J. Cardoso, and J. S. Cardoso. Inbreast: toward a full-field digital mammographic database. Academic radiology, 19(2):236–248, 2012.

32. D. Nie, R. Trullo, J. Lian, L. Wang, C. Petitjean, S. Ruan, Q. Wang, and D. Shen. Medical image synthesis with deep convolutional adversarial networks. IEEE Transactions on Biomedical Engineering, pages 1–1, 2018.

33. A. Odena, V. Dumoulin, and C. Olah. Deconvolution and checkerboard artifacts. Distill, 2016.

34. A. Osokin, A. Chessel, R. E. C. Salas, and F. Vaggi. Gans for biological image synthesis. ICCV, 2017.

35. X. Peng, Z. Tang, F. Yang, R. S. Feris, and D. Metaxas. 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, pages 2226–2234, 2018.

36. O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In MICCAI, 2015.

37. H. R. Roth, A. Farag, E. B. Turkbey, L. Lu, J. Liu, and R. M. Summers. Data from pancreas-ct. the cancer imaging archive. 2016.

38. H. R. Roth, L. Lu, N. Lay, A. P. Harrison, A. Farag, A. Sohn, and R. M. Summers. Spatial aggregation of holistically-nested convolutional neural networks for automated pancreas localization and segmentation. Medical Image Analysis, 45:94 – 107, 2018.

39. H. R. Roth, L. Lu, A. Seff, K. M. Cherry, J. Hoffman, S. Wang, J. Liu, E. Turkbey, and R. M. Summers. A new 2.5 d representation for lymph node detection using random sets of deep convolutional neural network observations. In MICCAI, pages 520–527, 2014.

40. T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. In NIPS, 2016.

41. A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb. Learning from simulated and unsupervised images through adversarial training. CVPR, 2017.

42. D. Ulyanov, A. Vedaldi, and V. Lempitsky. Instance normalization: The missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022, 2016.

43. H. Van Nguyen, K. Zhou, and R. Vemulapalli. Cross-domain synthesis of medical images using efficient location-sensitive deep network. In MICCAI, 2015.

44. R. Vemulapalli, H. Van Nguyen, and S. Kevin Zhou. Unsupervised cross-modal synthesis of subject-specific scans. In ICCV, 2015.

45. Y. Xue, T. Xu, H. Zhang, R. Long, and X. Huang. Segan: Adversarial network with multi-scale l_1 loss for medical image segmentation. arXiv preprint arXiv:1706.01805, 2017.

46. D. Yang, D. Xu, S. K. Zhou, B. Georgescu, M. Chen, S. Grbic, D. Metaxas, and D. Comaniciu. Automatic liver segmentation using an adversarial image-to-image network. MICCAI, 2017.

47. Z. Yi, H. Zhang, P. T. Gong, et al. Dualgan: Unsupervised dual learning for image-to-image translation. ICCV, 2017.

48. J. Zbontar, F. Knoll, A. Sriram, M. J. Muckley, M. Bruno, A. Defazio, M. Parente, K. J. Geras, J. Katsnelson, H. Chandarana, et al. fastmri: An open dataset and benchmarks for accelerated mri. arXiv preprint arXiv:1811.08839, 2018.

49. R. Zhang, P. Isola, and A. A. Efros. Colorful image colorization. In ECCV, 2016.

50. Z. Zhang, P. Chen, M. Sapkota, and L. Yang. Tandemnet: Distilling knowledge from medical images using diagnostic reports as optional semantic references. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 320–328, 2017.

51. Z. Zhang, Y. Xie, F. Xing, M. McGough, and L. Yang. Mdnet: A semantically and visually interpretable medical image diagnosis network. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 6428–6436, 2017.

52. Z. Zhang, Y. Xie, and L. Yang. Photographic text-to-image synthesis with a hierarchically-nested adversarial network. In CVPR, 2018.

53. Z. Zhang, L. Yang, and Y. Zheng. Translating and segmenting multimodal medical volumes with cycle- and shape-consistency generative adversarial network. In CVPR, 2018.

54. H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In CVPR, pages 6230–6239, 2017.

55. L. Zhao, X. Peng, M. Kapadia, and D. N. Metaxas. Learning to forecast and refine residual motion for image-to-video generation. In European conference on computer vision, 2018.

56. J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. ICCV, 2017.

57. Z. Zhu, Y. Xia, W. Shen, E. K. Fishman, and A. L. Yuille. A 3d coarse-to-fine framework for automatic pancreas segmentation. CoRR, abs/1712.00201, 2017.