ECAI 2023
K. Gal et al. (Eds.)

doi:10.3233/FAIA230633

Deep Co-Training for Cross-Modality Medical Image Segmentation

Lei Zhu[a;*], **Ling Ling Chan**[b], **Teck Khim Ng**[c], **Meihui Zhang**[d] **and Beng Chin Ooi**[c]

[a]Institute of High Performance Computing (IHPC), Agency for Science, Technology and Research (A*STAR), Singapore
[b]Singapore General Hospital, Singapore
[c]National University of Singapore, Singapore
[d]Beijing Institute of Technology, Beijing, China

Abstract. Due to the expensive segmentation annotation cost, cross-modality medical image segmentation aims to leverage annotations from a source modality (e.g. MRI) to learn a model for target modality (e.g. CT). In this paper, we present a novel method to tackle cross-modality medical image segmentation as semi-supervised multi-modal learning with image translation, which learns better feature representations and is more robust to source annotation scarcity. For semi-supervised multi-modal learning, we develop a deep co-training framework. We address the challenges of co-training on divergent labeled and unlabeled data distributions with a theoretical analysis on multi-view adaptation and propose decomposed multi-view adaptation, which shows better performance than a naive adaptation method on concatenated multi-view features. We further formulate inter-view regularization to alleviate overfitting in deep networks, which regularizes deep co-training networks to be compatible with the underlying data distribution. We perform extensive experiments to evaluate our framework. Our framework significantly outperforms state-of-the-art domain adaptation methods on three segmentation datasets, including two public datasets on cross-modality cardiac substructure segmentation and abdominal multi-organ segmentation and one large scale private dataset on cross-modality brain tissue segmentation. **Our code is publicly available at https://github.com/zlheui/DCT.**

1 Introduction

Deep learning has achieved great success in medical image analysis [17], however, it requires huge amount of labeled data to be effective, which is both expensive and time consuming to obtain. It is desirable to develop deep learning methods that are annotation-efficient. To this end, cross-modality medical image segmentation aims to leverage existing annotations from a source modality (e.g. MRI) to learn a model for target modality (e.g. CT). However, deep learning models trained in one modality usually give poor performance in another modality due to distribution shift. Unsupervised domain adaptation (UDA) methods have shown promising performance for cross-modality medical image segmentation task. State-of-the-art UDA methods adopt synergistic image and feature adaptation to reduce the distribution shift across domains at both image and feature level [3, 33, 10, 12]. However, these UDA methods may be sub-optimal, as they merely align the target data distribution with annotated source data without considering learning target data structure.

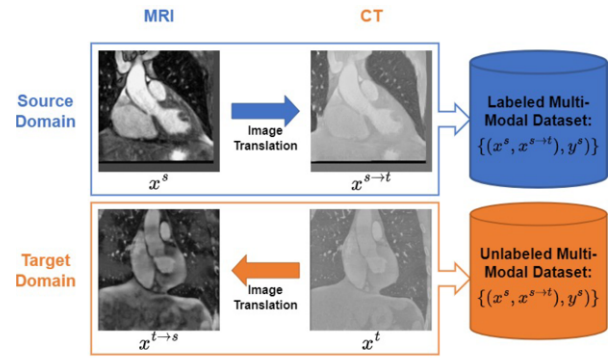

Figure 1. Image translation enables semi-supervised multi-modal learning. We adopt CycleGAN to translate source MRI image into CT image and target CT image into MRI image. The image translation results are quite good where the appearance of the original images are translated into the appearance of the other modality and the contents of the images are well-preserved. Augmenting the original single-modal datasets with the translated images creates one labeled multi-modal dataset in source domain and one unlabeled multi-modal dataset in target domain, which suggests semi-supervised multi-modal learning.

In this paper, we present a novel method to tackle cross-modality medical image segmentation with semi-supervised multi-modal learning. From Fig. 1, we can observe that existing image translation techniques can transform an MRI image into CT appearance and a CT image into MRI appearance while preserving the image content with fairly good quality [31]. Similarly, we should be able to augment the datasets in cross-modality medical image segmentation, namely, the labeled images in source modality (e.g. MRI) and unlabeled images in target modality (e.g. CT) with their translated images. In this way, we can obtain a labeled multi-modal dataset from source domain and an unlabeled multi-modal dataset from target domain as shown in figure.

As opposed to solving the cross-modality medical image segmentation task with domain adaptation, which merely aligns the target data distribution towards the source data distribution, we propose a semi-supervised multi-modal learning approach. Consequently, our goal is to learn a model that can perform well with complementary multi-modal information in a semi-supervised manner, which can

* Corresponding Author. Email: zhu_lei@ihpc.a-star.edu.sg

1). lead to better feature representations for both domains, as both labeled and unlabeled data are leveraged for learning discriminative deep feature representations; **2). be more robust to source annotation scarcity**, as a solution to semi-supervised multi-modal learning naturally handles the case when we have limited annotations in source domain. In essence, we propose to transform the task of cross-modality medical image segmentation into the task of semi-supervised multi-modal learning with image translation. Solving the latter task can potentially provide a better solution to the former task and is robust to annotation scarcity.

Co-training [1] is a semi-supervised multi-modal learning method, where two models are first learned on the two different views (modalities) of the labeled data. Subsequently, unlabeled data with model assigned pseudo-labels are gradually added to the labeled data set for continual training. We can apply co-training on our augmented datasets to learn two segmentation networks for the two different modalities. However, plain co-training with deep networks is unlikely to work. The challenges originate from both the dataset setting and the engagement of deep networks: 1). The augmented multi-modal datasets are synthesized from the labeled source dataset and unlabeled target dataset, which are drawn from different distributions. As can be observed in Fig. 1, despite the modality difference, there are still some morphological and scale differences between source and target data, which may be due to different patient distributions across domains or different machine scanning parameters; 2). Deep networks are notoriously known to require large scale labeled data to be effective and tend to overfit, which will deteriorate co-training performance.

To facilitate effective deep co-training, we address the two challenges in our framework as follows: First, to reduce the distribution shift between the source and target data, we conduct theoretical analysis on multi-view adaptation and develop a theorem to enable decomposition of adaptation with multi-view data into adaptation on each single view. Based on which, we propose decomposed multi-view adaptation which yields better performance than a naive adaptation method on concatenated multi-view features that does not consider the view-wise features. Second, co-training assumes target concepts to be compatible with the underlying data distribution to be effective. This can however be violated when deep networks overfit the labeled data and cannot generalize to unlabeled data. To this end, we introduce inter-view regularization to enforce consistency of predictions on different views of the same data point.

In summary, we have made following contributions in this paper:

- We develop a deep co-training framework for cross-modality medical image segmentation. Compared to existing UDA methods, our method learns better feature representations and is more robust to source annotation scarcity.
- We prove a theorem for multi-view adaptation and propose a general decomposed multi-view adaptation method, which shows better performance compared to adaptation with concatenated multi-view features.
- We propose inter-view regularization to regularize deep co-training networks to be compatible with synthesized multi-modal data, which is generally applicable for deep co-training.
- We conduct extensive experiments to evaluate our framework, where we have collected and processed a large scale private brain tissue segmentation dataset to verify that our framework can be effectively applied in real clinical settings. Our framework outperforms state-of-the-art cross-modality medical image segmentation methods significantly in all the three datasets we evaluate.

2 Related Works

Unsupervised Domain Adaptation has shown promising performance in cross-modality medical image segmentation task. Existing UDA methods can be mostly categorized into feature adaptation methods, image adaptation methods, and hybrid methods which combine feature and image adaptation. Feature adaptation methods reduce feature distribution shift by either minimizing certain distribution metric like maximum mean discrepancy (MMD) [30, 18], or through adversarial training with a domain discriminator [8, 29, 7]. Image adaptation methods [2, 13] translate image appearance across domains and learn a target model on translated source images. Hoffman et al. [11] are among the first to combine feature adaptation with image adaptation while enforcing semantic consistency with a static source trained model. Chen et al. [3] proposes synergistic feature and image adaptation which fuses part of image translation pipeline with feature representation learning. Zou et al. [33] introduces a target-to-source adaptation branch with a dual-scheme fusion network for more effective adaptation. Han et al. [10] proposes deep symmetric adaptation network with a bidirectional adaptation structure. Most recently, Hu et al. [12] introduce a semantic similarity constraint with contrastive learning [6] to further boost the cross-modality medical image segmentation performance. Unlike previous methods, self-training based UDA method CBST [34] use source trained network to assign pseudo-labels to unlabeled target data and then use pseudo-labeled target data to update the network for target data structure learning. Co-training for Domain Adaptation (CODA) [4] solves an optimization problem which simultaneously learns a target classifier, a split of the feature space into two different views, and a subset of source and target features. Their method is however limited to simple linear models and text-based classification tasks. Asymmetric Tri-training for Domain Adaptation (ATDA) [24] uses two source classifiers to assign pseudo labels to unlabeled target data and then uses the pseudo-labeled target data to train a target classifier with shared encoder network, but their method is limited to simple digit and review classification tasks. Most existing UDA methods focus on reducing the distribution shift across domains for knowledge transfer, thus, they are well-suited for cross-modality medical image segmentation task. However, unlike existing UDA methods, our deep co-training framework tackle cross-modality medical image segmentation as a semi-supervised multi-modal learning task via image translation.

Image Translation aims to translate images from one style into another style while preserving the image content. Most image translation methods are based on generative adversarial networks (GAN) framework [9]. DCGAN [23] proposes deep convolutional GAN architecture to learn better feature representations and improve the image translation quality. CycleGAN [31] introduces a cycle-consistency loss, which demonstrates great image translation results while preserving the image content. Image translation has already demonstrated good performance in UDA [2, 13] to translate source image into target style for learning. As far as we are concerned, we are the first to explore usage of image translation to synthesize multi-modal data for semi-supervised multi-modal learning.

Co-Training [1] is a method for semi-supervised multi-modal learning, where two models are trained on two modalities for learning complementary information. Co-training has been applied in various machine learning tasks like text classification [21], object recognition [5], domain adaptation [4], and deep semi-supervised classification [22]. As far as we are concerned, we are the first to investigate a deep co-training framework on synthesized multi-modal data for semi-supervised multi-modal segmentation.

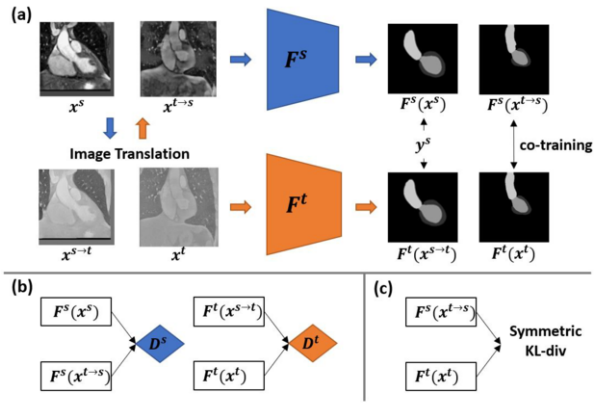

Figure 2. Overview of our proposed deep co-training framework. (a). Plain deep co-training on synthesized multi-modal data with image translation. There are two segmentation networks to perform segmentation in each modality. Labeled source data and unlabeled target data are utilized to perform co-training on the two networks. The plain deep co-training framework is unlikely to work due to distribution discrepancy and potential overfit of deep networks on limited annotations. Thus, we introduce the following two components into our framework: (b). Decomposed multi-view adaptation. Two domain discriminators are introduced into the framework to perform decomposed adaptation on each view separately based on our theorem; (c). Inter-view regularization is introduced to regularize the deep co-training networks to be compatible with the synthesized multi-modal target data to alleviate overfitting. Our whole framework is composed of the components in (a), (b), and (c).

3 Methods

In cross-modality medical image segmentation task, we are given N^s labeled data $\mathbb{D}^s = \{(\boldsymbol{x}_i^s, y_i^s)\}_{i=1}^{N^s}$ in source domain and N^t unlabeled data $\mathbb{D}^t = \{\boldsymbol{x}_i^t\}_{i=1}^{N^t}$ in target domain. The source and target data share the same set of C labels and are sampled from probability distributions P^s and P^t respectively with $P^s \neq P^t$. The goal is to learn a model with labeled source data and unlabeled target data that can perform well in target domain. In Fig. 2, we present an overview of our proposed deep co-training framework which tackles the task as semi-supervised multi-modal learning.

3.1 Image Translation and Deep Co-Training

We adopt CycleGAN [31] for image translation in our framework due to its good performance, while better image translation techniques will further boost the performance of our framework. With image translation, we augment the original dataset with their translated images. Denote $\widetilde{\mathbb{D}}^s = \{((\boldsymbol{x}_i^s, \boldsymbol{x}_i^{s\to t}), y_i^s)\}_{i=1}^{N^s}$ and $\widetilde{\mathbb{D}}^t = \{(\boldsymbol{x}_i^{t\to s}, \boldsymbol{x}_i^t)\}_{i=1}^{N^t}$ as the augmented source and target dataset respectively, where $\boldsymbol{x}_i^{s\to t}$ is the translated image of $\boldsymbol{x}_i^s$ in target modality and $\boldsymbol{x}_i^{t\to s}$ is the translated image of $\boldsymbol{x}_i^t$ in source modality.

In our deep co-training framework, we first learn two segmentation networks on the two different modalities with the following hybrid loss:

$$\mathcal{L}_{seg}^s = \mathbb{E}[H(y^s, F^s(\boldsymbol{x}^s)) + Dice(y^s, F^s(\boldsymbol{x}^s))], \quad (1)$$

$$\mathcal{L}_{seg}^{s\to t} = \mathbb{E}[H(y^s, F^t(\boldsymbol{x}^{s\to t})) + Dice(y^s, F^t(\boldsymbol{x}^{s\to t}))], \quad (2)$$

where F^s and F^t are two segmentation networks for source and target modality respectively, $H(\cdot)$ is the pixel-wise cross-entropy loss, which we assign class weights to balance different classes and $Dice(\cdot)$ is the widely adopted dice loss [19] . The hybrid loss is designed to sufficiently learn the two segmentation networks with complementary supervision signals.

Next, we perform co-training on the unlabeled target data. We extend class-balanced self-training [34] for deep co-training where the selected target pixels to label is the union of the selected target pixels from the two segmentation networks. The selection function S is defined as follows:

$$S(p_t) = \mathbb{1}_{[c=argmax_c p_t^{(c)} \wedge p_t^{(c)} > exp(-k_c)]}(p_t^{(c)}), \quad (3)$$

where p_t is the prediction mask, $\mathbb{1}$ is the indicator function which returns 1 if the condition is true and 0 otherwise and k_c is the class-balanced weights [34]. The final labeled target pixel is $\mathcal{S}(\boldsymbol{x}^t) = S(F^s(\boldsymbol{x}^{t\to s})) \cup S(F^t(\boldsymbol{x}^t))$. The co-training loss is defined as follows:

$$\mathcal{L}_{cot}^t = \mathbb{E}[H(\mathcal{S}(\boldsymbol{x}^t), F^t(\boldsymbol{x}^t))], \quad (4)$$

$$\mathcal{L}_{cot}^{t\to s} = \mathbb{E}[H(\mathcal{S}(\boldsymbol{x}^t), F^s(\boldsymbol{x}^{t\to s}))]. \quad (5)$$

Discussion. **Will the above plain deep co-training framework work?** As we mentioned before, it is challenging to use synthesized multi-modal dataset for deep co-training. Thus, the above framework is unlikely to work as effective. To this end, we introduce two extra components into our framework to ensure deep co-training works effectively. **Will back-propagating the supervision signal to train the image translation model help improve the performance?** Currently, the image translation model and the segmentation networks are trained in isolation. However, as the segmentation networks receive the translated images as input and possess semantic knowledge on each class, we think back-propagating the supervision signal from the segmentation networks to train image translation model from end-to-end would help boost the performance of our framework. We provide empirical results in Sec. 4.3.

3.2 Decomposed Multi-View Adaptation

In our problem, source and target data are drawn from different data distributions. As observed in Fig. 1, there is difference between the source image pair and target image pair despite the modality difference. Thus, it is necessary to reduce the distribution shift between the labeled source data and unlabeled target data. One naive solution is to concatenate multi-view features and use a domain discriminator to discriminate the concatenated features from source and target data for distribution alignment without considering the view-wise features. However, using a single domain discriminator may fail to capture the minor differences for features within a single view. We develop a theorem which states that we can decompose multi-view adaptation into adaptation in each single view:

Theorem 1 *Let $\mathcal{H}$ be a hypothesis space of VC dimension d and let P^s and P^t be the data distribution for source data and target data respectively. Suppose data instances in both source distribution and target distribution have k different views, where $\boldsymbol{x}^s = (\boldsymbol{v}_1^s, \boldsymbol{v}_2^s, ..., \boldsymbol{v}_k^s)$ for each $(\boldsymbol{x}^s, y^s) \in P^s$. Similarly for $(\boldsymbol{x}^t, y^t) \in P^t$. If U^s, U^t are unlabeled data of size m' each drawn from P^s and P^t respectively, then for any $\delta \in (0, 1)$, with probability at least 1-δ, for*

every $h_1, h_2, ..., h_k \in \mathcal{H}$, we have:

$$\epsilon_t(h) \leq \frac{1}{k}\sum_{i=1}^{k}\left(\epsilon_s(h_i) + \frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(U_i^s, U_i^t) + C_i\right) + 4\sqrt{\frac{2dlog(2m') + log(\frac{2}{\delta})}{m'}}, \tag{6}$$

where $\epsilon_s(\cdot)$ (resp. $\epsilon_t(\cdot)$) measures the expectation error of a hypothesis on source (resp. target) data distribution, $h = \frac{\sum_{i=1}^{k} h_i}{k}$ is the composite hypothesis, $d_{\mathcal{H}\Delta\mathcal{H}}(U_i^s, U_i^t)$ is the empirical estimation of the $\mathcal{H}\Delta\mathcal{H}$-distance on the i-th view of unlabeled data U^s and U^t, $C_i = \min_{h_i \in \mathcal{H}} \epsilon_t(h_i) + \epsilon_s(h_i)$.

The above theorem states that for the composite multi-view model, its performance on target distribution is upper bounded by the performance of each constituent model on source distribution and the distribution shift between source and target distribution in each view plus some constant terms. In deep co-training, the model performance on target distribution affects the accuracy of the assigned labels for unlabeled target data. Thus, the theorem confirms the necessity in minimizing the distribution shift across source and target data. In addition, the theorem states that it suffices to reduce the distribution shift for each view separately. Based on our theorem, we proposed decomposed multi-view adaptation. Specifically, we introduce two domain discriminators into our framework, namely D^s and D^t to separately reduce the distribution shift in the two different modalities and we control the strength of adaptation with a balancing weight. We use D^s to discriminate prediction masks from source data and translated target data, and D^t to discriminate prediction masks from translated source data and target data. We train the two segmentation networks F^s and F^t adversarially so that the learned features become domain invariant to produce similar prediction masks across domains. The adversarial training losses are defined as follows:

$$\mathcal{L}_{adv}^s = \mathbb{E}[logD^s(F^s(\boldsymbol{x}^s))] + \mathbb{E}[log(1 - D^s(F^s(\boldsymbol{x}^{t\to s})))], \tag{7}$$

$$\mathcal{L}_{adv}^t = \mathbb{E}[logD^t(F^t(\boldsymbol{x}^{s\to t}))] + \mathbb{E}[log(1 - D^t(F^t(\boldsymbol{x}^t)))]. \tag{8}$$

3.3 Inter-View Regularization

The success of co-training relies on the "compatibility" assumption among the target concepts in each view and the underlying data distribution [1], namely if F^* denotes the combined target concept and F_1^* and F_2^* denote the target concept in each view, then for any example $\boldsymbol{x} = (\boldsymbol{v}_1, \boldsymbol{v}_2)$, we have $F^*(\boldsymbol{x}) = F_1^*(\boldsymbol{v}_1) = F_2^*(\boldsymbol{v}_2)$. Intuitively, the "compatibility" assumption enables us to use model from one view to assign labels to unlabeled data and then use the other view of unlabeled data with assigned labels to train the other model, which is at the core of co-training.

In the original algorithm [1], models are trained on labeled data and perform co-training on unlabeled data without regularization. This is because the original models are simple linear models and regularization is not needed. However, deep learning models are representation learning, have high capacities, and are easy to overfit. Moreover, the labeled source data can be scarce. Consequently, the "compatibility" assumption can be violated when models overfit the labeled data and cannot generalize to unlabeled data. Thus, it is necessary to regularize the two segmentation networks in our deep co-training framework to conform to the "compatibility" assumption. To this end, we propose inter-view regularization with synthesized multi-modal data to ensure the predictions on the original and translated data to be compatible. Specifically, we input the target data and the translated target data into the corresponding segmentation network to obtain their prediction masks. Then, we minimize the discrepancy for the predicted probability vectors at each pixel to ensure the two segmentation network have similar predictions. We choose symmetric Kullback-Leibler (KL)-divergence which measures how one probability distribution is different from another as follows:

$$\mathcal{L}_{reg} = \mathbb{E}[\frac{1}{2}(KL(F^t(\boldsymbol{x}^t), F^s(\boldsymbol{x}^{t\to s})) + KL(F^s(\boldsymbol{x}^{t\to s}), F^t(\boldsymbol{x}^t)))], \tag{9}$$

where $KL(\cdot,\cdot)$ measures the average pixel-wise KL-divergence between two prediction masks.

Discussion. Consistency regularization is widely adopted in semi-supervised learning to regularize the learning of deep networks to avoid overfitting. They usually input two perturbed data points into the same network and ensure the network to make similar predictions [25, 27]. Some use mean teacher [28], some use virtual adversarial training [20], and some use the same input into two different networks [22] for regularization. Different from them, we are the first to use synthesized multi-modal data for regularization; the regularization method is proposed to enable co-training with deep networks; and we focus on segmentation task as opposed to classification.

Training objective: The overall objective function of our deep co-training framework is as follows:

$$\mathcal{L}_{all} = \mathcal{L}_{seg}^s + \mathcal{L}_{seg}^{s\to t} + \mathcal{L}_{cot}^t + \mathcal{L}_{cot}^{t\to s} + \lambda_{adv}(\mathcal{L}_{adv}^s + \mathcal{L}_{adv}^t) + \lambda_{reg}\mathcal{L}_{reg}, \tag{10}$$

where λ_{adv} and λ_{reg} are the balancing weights, which are both set to 1 empirically.

4 Experiments

4.1 Datasets and Evaluation Metrics

We validate the effectiveness of our framework with three datasets: cardiac substructure segmentation [32]; abdominal multi-organ segmentation [15, 16]; and **a large-scale private brain tissue segmentation dataset.** More details about the large scale private brain tissue segmentation dataset can be found in our supplementary material.

The cardiac dataset consists of 20 unpaired MRI and CT volumes with ground truth masks on four heart substructures: ascending aorta (AA), left atrium blood cavity (LAC), left ventricle blood cavity (LVC), and myocardium of the left ventricle (MYO). The abdominal dataset consists of 20 unpaired T2-SPIR MRI and 30 CT volumes collected from two public datasets with ground truth masks on four organs: spleen, right kidney, left kidney, and liver. The private brain tissue segmentation dataset consists of 968 paired MRI and CT volumes with ground truth masks on three types of tissues: cerebrospinal fluid (CSF), grey matter (GM), and white matter (WM). All the data are cropped, normalized with zero mean and unit variance, and resampled into the size of 256×256. Coronal view of cardiac volumes and axial view of abdominal volumes and brain volumes are used to train the 2D network. Both MRI to CT and CT to MRI transfer are considered for the two public datasets. MRI to CT transfer is considered for the private dataset. Each modality is randomly split with 80% scans for training and 20% for testing following existing studies [3, 33, 12].

We employ three commonly-used metrics, namely the Dice similarity coefficient (Dice), continuous Dice similarity coefficient

Table 1. Ablation study of our deep co-training framework on Abdominal Multi-Organ MRI→CT transfer task. DCT = Deep Co-Training, DMA = Decomposed Multi-view Adaptation, IVR = Inter-View Regularization, DCT-SEP = Deep Co-Training without back-propagating gradients to image translation model, DCT-CA = Deep Co-Training with adaptation on concatenated multi-view features, DCT-CA↑ = DCT-CA with increased domain discriminator size.

Abdominal Multi-Organ Segmentation Performance MRI→CT								
Method	$\mathcal{L}_{seg}$	$\mathcal{L}_{cot}$	$\mathcal{L}_{adv}$	$\mathcal{L}_{reg}$	Dice			ASD
					F^s	F^t	Ensemble	Ensemble
Source only	✓				NA	NA	62.0	4.3
Plain DCT	✓	✓			32.3	41.8	35.3	15.4
Plain DCT + DMA	✓	✓	✓		58.4	67.4	74.4	7.1
Plain DCT + IVR	✓	✓		✓	85.3	85.7	86.4	2.1
DCT (Our Proposed)	✓	✓	✓	✓	**87.1**	**87.8**	**88.0**	**1.6**
DCT-SEP	✓	✓	✓	✓	81.8	84.9	85.0	2.2
DCT-CA	✓	✓	✓	✓	85.0	85.1	85.7	2.3
DCT-CA↑	✓	✓	✓	✓	84.0	87.0	87.2	2.3

(cDice), and the average symmetric surface distance (ASD) to quantitatively evaluate the segmentation performance. Dice measures the voxel-wise segmentation accuracy between the predicted and reference volumes. cDice [26] is a variant of the Dice coefficient that evaluate spatial similarity between binary images and real-valued probability maps. ASD calculates the average distances between the surface of the prediction mask and the ground truth in 3D. A higher Dice and cDice value or a lower ASD value indicates better segmentation results. The evaluation is performed on the subject-level segmentation volume.

4.2 Implementation Details

For the implementation of our deep co-training framework, the image translation network is implemented and trained according to the original CycleGAN paper [31]. We implement the segmentation networks in our framework follow the same architecture as [3, 33] for fair comparison, which consists of twelve convolutional operation groups, two dilated convolutional groups and one softmax layer. The domain discriminator networks in our framework is implemented following the architecture of PatchGAN [14], which have five convolutional layers with channels size of 64, 128, 256, 512, and 1, respectively. We use Adam optimizer with learning rate of 2×10^{-4}. We split the training of our framework into two phases. In the first phase, we train our framework without co-training loss $\mathcal{L}_{cot}$ for $10k$ iterations to warm up the segmentation networks. In the second phase, we include the co-training loss and train the framework for another $10k$ iterations. The batch size is set to 6 on a NVIDIA GeForce GTX 1080p GPU. For final prediction, we ensemble the predictions from both segmentation networks by averaging their prediction probabilities. We run all experiments three times and report the mean.

4.3 Ablation Study

We perform extensive ablation studies to investigate how our designs contribute to a deep co-training framework for cross-modality medical image segmentation. Table 1 shows the experiment results. First, the plain deep co-training framework fails to work due to the source and target data distribution discrepancy and the violation of co-training "compatibility" assumption when deep networks overfit. Second, the addition of either our proposed decomposed multi-view adaptation or inter-view regularization technique tackles one of the above two challenges and helps to boost the performance of our framework to be better than the source only baseline. Third, the combination of both components into our framework achieves the best performance, which is due to the complementary roles of them played in enabling deep co-training.

Next, we present ablation studies on some other aspects of our framework. First, our ablation studies show that our framework can gain about 3 points boost in dice score when we back-propagate the supervision signal from the segmentation networks to train the image translation model compared to when we do not as shown in Table 1. Second, for the final prediction, we ensemble the predictions from the two segmentation networks. Our ablation studies show that ensemble generally helps to improve the results when compared to using either single segmentation network for final prediction.

Finally, we have proposed decomposed multi-view adaptation, which is a general methodology for multi-view adaptation. We compare it with a naive adaptation method on concatenated multi-view features. The results show that decomposed multi-view adaptation leads to better results compared to adaptation on concatenated multi-view features. To ensure the difference is not due to the larger capacity of our method as our method have two domain discriminators, we double the size of the domain discriminator in the comparison method. We find that increasing the size of domain discriminator helps to improve the performance in the comparison method, however, our method still outperforms it. The experiment results indicate that dedicated adaptation for each single view is better than adaptation on the concatenated multi-view features without considering the view-wise features, where adaptation on the concatenated multi-view features may fail to differentiate the minor differences in each single view.

4.4 Comparison with State-of-The-Art

We compare our deep co-training framework with state-of-the-art UDA methods for cross-modality medical image segmentation including CBST [34], ATDA [24], SynSeg-Net [13], CycleGAN [31], PnP-AdaNet [7], AdaOutput [29], CyCADA [11], DSFN [33], SIFA-v2 [3], DSAN [10], and SSC [12]. CBST is self-training based UDA method. ATDA is tri-training based UDA method. SynSeg-Net and CycleGAN are image adaptation based UDA methods. PnP-AdaNet and AdaOutput are feature adaptation based UDA methods. CyCADA, DSFN, SIFA-v2, DSAN, and SSC are joint image and feature adaptation UDA methods. In particular, SIFA-v2, DSFN, DSAN, and SSC all perform synergistic image and feature adaptation and are designed for medical image analysis. To demonstrate the domain shift across domains, we present the performance lower bound "Source only" by directly applying the model trained in source domain to target data. We also provide the performance upper bound "Supervised training" by training the model on target labels.

Table 2 presents both the experiment results for cardiac substructure segmentation and abdominal multi-organ segmentation. As can be observed, our deep co-training framework significantly outperform state-of-the-art UDA methods for cross-modality medical image segmentation. Specifically, for the Cardiac MRI→CT transfer task, our deep co-training framework has improved the average result by 1.0 point in Dice score compared to the previously best method. And for the more challenging Cardiac CT→MRI transfer task, our framework has improved the average result by 8.2 points in Dice score and reduced the ASD score by 1.6 points compared to the previously best method. For abdominal multi-organ segmentation, the improvement of our deep co-training framework upon SIFA-v2 outperforms state-of-the-art UDA method SSC by 2.0 points in dice score and 0.2 points in ASD score for MRI→CT transfer task. For

Table 2. Performance comparison with state-of-the-art domain adaptation methods on cardiac substructure segmentation and abdominal multi-organ segmentation. Numbers before the slash '/' are for MRI to CT transfer, after the slash '/' are for CT to MRI transfer. Results for method with * are cited from their paper. '+' and '-' denotes the increment or decrement upon SIFA-v2. **Bold** number highlights the best performance or best improvement upon SIFA-v2. Note that we compare our method with SSC and DSAN by improvement upon SIFA-v2 as both codes for SSC and DSAN are not publicly available and we preprocess the multi-organ dataset differently compared to them.

Cardiac Substructure Segmentation Performance (MRI→CT / CT→MRI)										
Method	Dice					ASD				
	AA	LAC	LVC	MYO	Avg	AA	LAC	LVC	MYO	Avg
Supervised training	83.2/82.8	90.5/86.5	92.0/92.4	88.3/79.1	88.5/85.2	2.3/3.8	2.3/2.1	1.7/2.0	1.5/1.6	1.9/2.3
Source only	11.4/0.8	40.3/21.3	8.7/30.4	0.4/10.9	15.2/15.8	33.9/24.7	29.3/19.6	34.3/10.9	34.8/7.6	33.1/15.7
CBST [34]	16.6/15.7	27.8/34.5	12.5/46.4	3.5/32.5	15.1/32.3	36.8/25.3	34.1/19.5	34.7/12.5	31.7/14.0	34.3/17.8
ATDA [24]	46.4/28.5	28.4/37.7	2.7/54.5	2.2/13.6	19.9/33.6	30.1/17.8	41.1/12.6	18.0/14.4	45.6/7.7	33.7/13.1
SynSeg-Net [13]	71.6/41.3	69.0/57.5	51.6/63.6	40.8/36.5	58.2/49.7	11.7/8.6	7.8/10.7	7.0/5.4	9.2/5.9	8.9/7.6
CycleGAN [31]	73.8/64.3	75.7/30.7	52.3/65.0	28.7/43.0	57.6/50.7	11.5/5.8	13.6/9.8	9.2/6.0	8.8/5.0	10.8/6.6
PnP-AdaNet [7]	74.0/43.7	68.9/47.0	61.9/77.7	50.8/48.6	63.9/54.3	12.8/11.4	6.3/14.5	17.4/4.5	14.7/5.3	12.8/8.9
AdaOutput [29]	65.2/60.8	76.6/39.8	54.4/71.5	43.6/35.5	59.9/51.9	17.9/5.7	5.5/8.0	5.9/4.6	8.9/4.6	9.6/5.7
CyCADA [11]	72.9/60.5	77.0/44.0	62.4/77.6	45.3/47.9	64.4/57.5	9.6/7.7	8.0/13.9	9.6/4.8	10.5/5.2	9.4/7.9
DSFN[33]	81.5/53.0	82.7/62.3	76.9/69.0	60.0/36.7	75.3/55.2	11.4/7.5	5.2/8.1	4.6/5.4	4.2/4.9	6.4/6.4
SIFA-v2 [3]	81.3/67.0	79.5/60.7	73.8/75.1	61.6/45.8	74.1/62.1	7.9/6.2	6.2/9.8	5.5/4.4	8.5/4.4	7.0/6.2
DSAN* [10]	79.9/71.3	84.8/66.2	82.8/76.2	66.5/52.1	78.5/66.5	7.7/4.4	6.7/7.3	3.8/5.5	5.6/4.3	5.9/5.4
SSC* [12]	82.0/NA	85.3/NA	**88.4**/NA	67.6/NA	80.8/NA	**6.2**/NA	4.1/NA	**3.0**/NA	**3.4**/NA	**4.2**/NA
Deep Co-Training (Our Proposed)	**86.7**/**72.6**	**85.5**/**75.7**	84.8/**87.2**	**70.5**/**63.4**	**81.8**/**74.7**	7.5/**5.2**	**3.2**/**4.2**	**3.0**/**2.6**	3.7/**3.4**	**4.2**/**3.8**
Abdominal Multi-Organ Segmentation Performance (MRI→CT / CT→MRI)										
Method	Dice					ASD				
	Spleen	R. kidney	L. kidney	Liver	Avg	Spleen	R. kidney	L. kidney	Liver	Avg
Supervised training	93.8/91.0	89.9/94.4	94.1/92.6	93.8/94.6	92.9/93.1	0.6/1.2	3.4/0.3	0.8/1.1	1.3/0.6	1.5/0.8
Source only	66.2/28.4	66.3/11.7	61.9/46.7	52.5/73.1	61.7/40.0	5.4/11.4	4.8/25.7	3.0/2.3	4.2/2.1	4.4/10.4
CBST [34]	81.8/81.5	77.3/81.8	85.0/86.5	83.4/77.9	81.9/81.9	6.0/2.9	4.4/1.1	2.8/1.9	3.8/2.9	4.3/2.2
ATDA [24]	85.5/43.0	67.7/3.7	62.2/48.6	77.7/30.8	73.3/31.5	3.8/7.8	7.7/24.0	15.9/7.4	7.9/10.7	8.8/12.5
SynSeg-Net [13]	81.1/85.3	82.6/83.9	82.8/87.0	83.8/83.5	82.6/84.9	1.9/1.5	2.4/0.9	2.5/0.9	4.5/2.5	2.8/1.5
CycleGAN [31]	83.3/79.4	80.7/84.4	82.9/89.1	87.4/87.4	83.6/85.1	2.2/2.3	2.8/1.0	1.7/0.7	2.4/2.2	2.3/1.5
AdaOutput [29]	87.2/80.0	81.7/87.1	86.0/85.2	84.0/85.5	84.7/84.5	1.6/0.8	3.2/**0.6**	1.6/0.8	2.3/1.5	2.2/0.9
CyCADA [11]	86.2/76.2	**84.8**/86.3	82.6/88.0	85.8/**90.3**	84.9/85.2	1.9/1.3	**2.0**/**0.6**	1.9/**0.6**	2.3/**1.0**	2.0/0.9
DSFN [33]	82.4/78.3	83.2/89.1	84.4/**90.7**	83.3/87.2	83.3/86.3	2.1/4.1	2.6/1.2	1.8/**0.6**	4.3/1.6	2.7/1.9
SIFA-v2 [3]	83.4/86.9	80.1/89.2	86.6/80.4	87.7/88.6	84.5/86.3	1.5/1.7	2.3/**0.6**	1.5/0.8	1.9/1.3	1.8/1.1
Deep Co-Training (Our Proposed)	**89.2**/**89.3**	81.7/**89.8**	**87.2**/87.3	**89.9**/86.2	**87.0**/**88.1**	**1.2**/**0.5**	2.4/0.7	**1.4**/**0.6**	**1.5**/1.4	**1.6**/**0.8**
DSAN* [10]	NA/-0.6	NA/**+2.3**	NA/+3.5	NA/**+2.3**	NA/**+1.8**	NA/+0.3	NA/**-0.1**	NA/**-0.4**	NA/**-0.5**	NA/-0.2
SSC* [12]	+0.5/NA	+0/NA	**+1.1**/NA	+0.5/NA	+0.5/NA	+0.1/NA	**+0**/NA	**-0.3**/NA	+0/NA	+0/NA
Deep Co-Training (Our Proposed)	**+5.8**/**+2.4**	**+1.6**/+0.6	+0.6/**+6.9**	**+2.2**/-2.4	**+2.5**/**+1.8**	**-0.3**/**-1.2**	+0.1/+0.1	-0.1/-0.2	**-0.4**/+0.1	**-0.2**/**-0.3**

Table 3. Performance comparison with state-of-the-art domain adaptation methods on brain tissue segmentation. **Bold** number highlights the best performance.

Brain Tissue Segmentation Performance MRI→CT								
Method	cDice				ASD			
	CSF	GM	WM	Avg	CSF	GM	WM	Avg
Supervised training	79.6	74.2	84.8	79.6	0.7	0.7	0.8	0.7
Source only	12.7	34.3	9.7	18.9	16.7	5.6	11.1	11.1
CBST [34]	33.4	50.5	37.0	40.3	13.0	6.4	17.4	12.3
SynSeg-Net [13]	66.0	57.7	15.9	46.5	1.3	**0.8**	2.7	1.6
AdaOutput [29]	60.0	60.6	23.9	48.2	1.5	0.9	5.0	2.5
SIFA-v2 [3]	67.1	60.7	53.9	60.6	1.2	0.9	1.7	1.2
Deep Co-Training (Our Proposed)	**75.8**	**66.1**	**75.9**	**72.6**	**1.1**	**0.8**	**1.4**	**1.1**

Table 4. Evaluation on the feature representations learned by the two constituent segmentation networks of our deep co-training framework and that of SIFA-v2 on cardiac substructure segmentation dataset. DCT-F^s = The F^s network in our deep co-training framework, DCT-F^t = The F^t network in our deep co-training framework.

Cardiac Substructure Segmentation Performance				
Method (Dice)	Source Representation		Target Representation	
	MRI→CT	CT→MRI	MRI→CT	CT→MRI
SIFA-v2 [3]	88.9	83.2	78.6	70.7
DCT-F^s	**91.0**	**84.5**	87.0	79.0
DCT-F^t	90.9	84.1	**87.2**	**80.2**

CT→MRI transfer task, the improvement of our framework upon SIFA-v2 is the same as DSAN in dice score and 0.1 points better in ASD score. Note that we compare our method with SSC and DSAN by improvement upon SIFA-v2 as both codes for SSC and DSAN are not publicly available and we preprocess the multi-organ dataset differently compared to them. Finally, the performance of our framework also approaches the supervised training upper bound.

Table 3 presents the experiment results on our large scale private brain tissue segmentation dataset. As can be observed, our framework outperforms state-of-the-art UDA methods significantly. Specifically, our framework has improved cDice score by 12.0 points and reduced ASD by 0.1 points compared to state-of-the-art UDA method SIFA-v2. Note that we do not compare with more advanced DSAN and SSC methods as their codes are not publicly available. But our experiment results on the two public datasets already demonstrate the effectiveness of our framework when compared to theirs.

Fig. 3 shows the visual comparison results for cardiac substructure segmentation. Due to space limit, we put more visual comparison results on abdominal multi-organ segmentation in the supplementary material. We do not present the visualization results on brain tissue segmentation as the dataset is private. As can be seen in the figure, the segmentation masks produced by our deep co-training framework are closer to the ground truth and contain fewer wrong semantic prediction results. However, as shown in the forth row in Fig. 3, all UDA methods fail to segment a small portion of ascending aorta, which is disconnected from the main part due to slice cut. But the supervised learning method can accurately segment that portion out, which indicates there is still a gap between existing UDA methods and supervised learning method that needs to be filled in future works.

4.5 *Discussion*

Deep Co-Training Learns Better Feature Representations for Both Domains. One of our arguments to tackle the cross-modality medical image segmentation task as semi-supervised multi-modal learning is that semi-supervised multi-modal learning can leverage

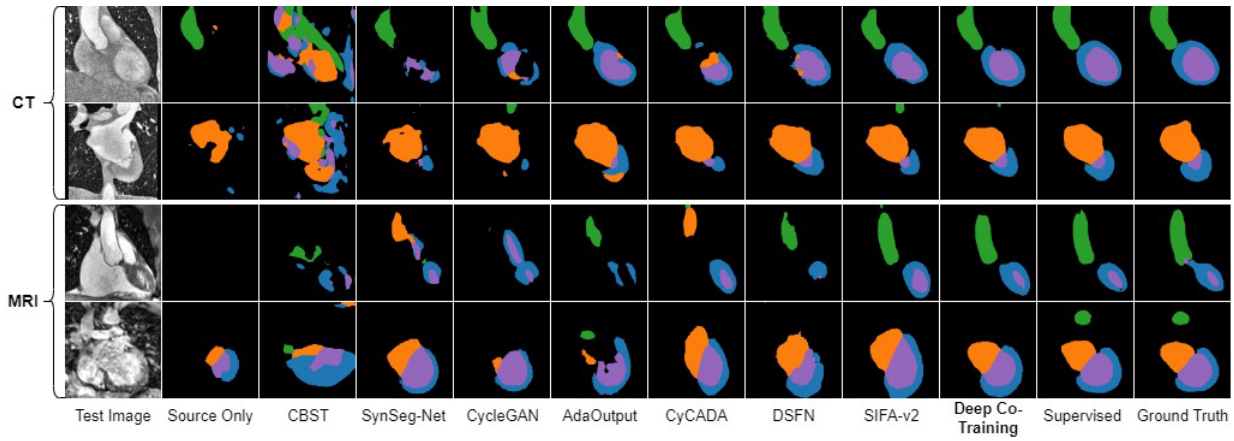

Figure 3. Visual comparison of segmentation results with different unsupervised domain adaptation methods for cardiac CT images and MRI images. The cardiac substructure of AA, LAC, LVC and MYO are indicated in green, orange, purple, blue colors respectively.

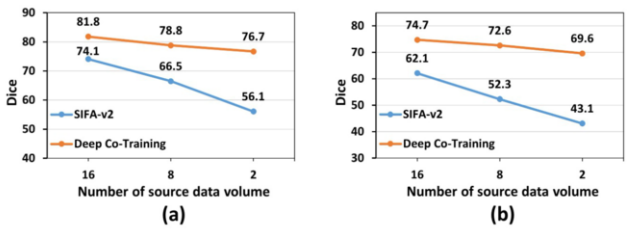

Figure 4. Evaluation on the performance of our framework with reduced source data annotations on cardiac substructure segmentation (a) MRI→CT transfer task and (b) CT→MRI transfer task.

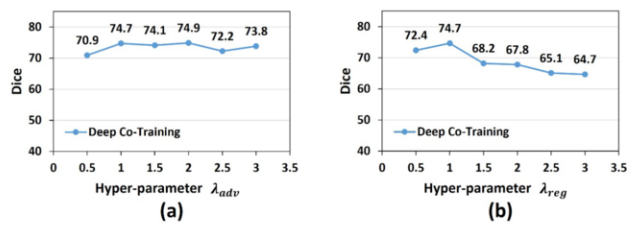

Figure 5. Sensitivity analysis of our framework on cardiac substructure segmentation CT→ MRI transfer task with (a) λ_{adv} and (b) λ_{reg}.

the complementary multi-modal information to learn better feature representations for both domains. To validate it, we evaluate the feature representations learned by the two constituent segmentation networks of our framework and that of state-of-the-art UDA method SIFA-v2 on cardiac substructure segmentation dataset. For source representation, we fix the feature learned by our framework and SIFA-v2, and fine tune the last layer of segmentation network on labeled source data and report performance on source test data. Similarly we do that for target representation. The experiment results are shown in Table 4. As can be seen, both the two constituent segmentation networks in our framework learns much better feature representations compared to that of SIFA-v2 in both domains due to the leverage of complementary multi-modal information in co-training.

Deep Co-Training is More Robust to Source Annotation Scarcity. As our framework tackles the cross modality medical image segmentation task as semi-supervised multi-modal learning, our framework naturally handles the case when we have limited annotations in source domain. To verify it, we compare our framework with the state-of-the-art UDA method SIFA-v2 [3] when we decrease the annotated data size in source domain. Fig. 4 shows the experiment results. As we can observe, for all source annotation sizes, our framework significantly outperforms SIFA-v2; with the decrease of the annotation size, the drop of performance in our framework is much smaller than SIFA-v2; more importantly, our framework with only 2 annotated source data volume outperforms SIFA-v2 with 16 annotated source data volume. The experiment results highlight the wide applicability of our framework even under extreme annotation scarcity scenarios.

Sensitivity Analysis. We perform post-experiment sensitivity analysis with the two balancing weights of our framework, namely λ_{adv} and λ_{reg}. As can be seen in Fig. 5, our framework is generally robust to the change of λ_{adv} in a wide range. For λ_{reg}, too large or too small regularization either hinders co-training or fails to effectively regularize the deep networks, which leads to poor performance. Yet even with the worst value of λ_{reg} in Fig. 5(b), our framework still performs better than SIFA-v2 and close to the previously best method DSAN. The empirical value of $\lambda_{reg} = 1$ gives the best performance.

5 Conclusions

In this paper, we propose a novel method to tackle cross-modality medical image segmentation via converting the task into semi-supervised multi-modal learning with image translation. To this end, we propose a deep co-training framework, where we address the challenges of co-training on divergent labeled and unlabeled data distributions with theoretical analysis on multi-view adaptation and propose decomposed multi-view adaptation, which is a general multi-view adaptation methodology and shows better performance than adaptation with concatenated multi-view features. We further formulate inter-view regularization to tackle the challenge of co-training with deep networks. Our inter-view regularization is a general regularization method to make deep co-training networks to be compatible with the underlying data distribution. We perform extensive experiments to evaluate our framework. We further evaluate our framework with a large scale private dataset to test its applicability in real clinical settings. Our framework significantly outperforms state-of-the-art UDA methods on all three segmentation tasks, learns better feature representations, and is more robust to source data scarcity.

Acknowledgements

We would like to thank the reviewers for their comments, which helped improve this paper considerably. This work was supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG2-TC-2021-003), Agency for Science, Technology and Research (A*STAR) through its AME Programmatic Funding Scheme Under Project A20H4b0141, and under its RIE2020 Health and Biomedical Sciences (HBMS) Industry Alignment Fund Pre-Positioning (IAF-PP) grant no. H20C6a0032.

References

[1] Avrim Blum and Tom Mitchell, 'Combining labeled and unlabeled data with co-training', in *Proceedings of the eleventh annual conference on Computational learning theory*, pp. 92–100, (1998).

[2] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan, 'Unsupervised pixel-level domain adaptation with generative adversarial networks', in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3722–3731, (2017).

[3] Cheng Chen, Qi Dou, Hao Chen, Jing Qin, and Pheng Ann Heng, 'Unsupervised bidirectional cross-modality adaptation via deeply synergistic image and feature alignment for medical image segmentation', *IEEE Transactions on Medical Imaging*, (2020).

[4] Minmin Chen, Kilian Q Weinberger, and John Blitzer, 'Co-training for domain adaptation', *Advances in neural information processing systems*, **24**, (2011).

[5] Minmin Chen, Kilian Q Weinberger, and Yixin Chen, 'Automatic feature decomposition for single view co-training', in *ICML*, (2011).

[6] Xinlei Chen and Kaiming He, 'Exploring simple siamese representation learning', in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15750–15758, (2021).

[7] Qi Dou, Cheng Ouyang, Cheng Chen, Hao Chen, Ben Glocker, Xiahai Zhuang, and Pheng-Ann Heng, 'Pnp-adanet: Plug-and-play adversarial domain adaptation network with a benchmark at cross-modality cardiac segmentation', *arXiv preprint arXiv:1812.07907*, (2018).

[8] Yaroslav Ganin and Victor Lempitsky, 'Unsupervised domain adaptation by backpropagation', in *International conference on machine learning*, pp. 1180–1189. PMLR, (2015).

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

[10] Xiaoting Han, Lei Qi, Qian Yu, Ziqi Zhou, Yefeng Zheng, Yinghuan Shi, and Yang Gao, 'Deep symmetric adaptation network for cross-modality medical image segmentation', *IEEE transactions on medical imaging*, **41**(1), 121–132, (2021).

[11] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell, 'Cycada: Cycle-consistent adversarial domain adaptation', in *International conference on machine learning*, pp. 1989–1998. PMLR, (2018).

[12] Tao Hu, Shiliang Sun, Jing Zhao, and Dongyu Shi, 'Enhancing unsupervised domain adaptation via semantic similarity constraint for medical image segmentation', in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, ed., Lud De Raedt, pp. 3071–3077. International Joint Conferences on Artificial Intelligence Organization, (7 2022). Main Track.

[13] Yuankai Huo, Zhoubing Xu, Hyeonsoo Moon, Shunxing Bao, Albert Assad, Tamara K Moyo, Michael R Savona, Richard G Abramson, and Bennett A Landman, 'Synseg-net: Synthetic segmentation without target modality ground truth', *IEEE transactions on medical imaging*, **38**(4), 1016–1025, (2018).

[14] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros, 'Image-to-image translation with conditional adversarial networks', in *IEEE Conference on Computer Vision and Pattern Recognition*, (2017).

[15] A Emre Kavur, N Sinem Gezer, Mustafa Barış, Sinem Aslan, Pierre-Henri Conze, Vladimir Groza, Duc Duy Pham, Soumick Chatterjee, Philipp Ernst, Savaş Özkan, et al., 'Chaos challenge-combined (ct-mr) healthy abdominal organ segmentation', *Medical Image Analysis*, **69**, 101950, (2021).

[16] B. Landman, Z. Xu, J. E. Iglesias, M. Styner, T. R. Langerak, and A. Klein, 'Multi-atlas labeling beyond the cranial vault', (2020).

[17] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez, 'A survey on deep learning in medical image analysis', *Medical image analysis*, **42**, 60–88, (2017).

[18] Mingsheng Long, Jianmin Wang, Guiguang Ding, Jiaguang Sun, and Philip S Yu, 'Transfer joint matching for unsupervised domain adaptation', in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1410–1417, (2014).

[19] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi, 'V-net: Fully convolutional neural networks for volumetric medical image segmentation', in *2016 fourth international conference on 3D vision (3DV)*, pp. 565–571. IEEE, (2016).

[20] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii, 'Virtual adversarial training: a regularization method for supervised and semi-supervised learning', *IEEE transactions on pattern analysis and machine intelligence*, **41**(8), 1979–1993, (2018).

[21] Kamal Nigam and Rayid Ghani, 'Analyzing the effectiveness and applicability of co-training', in *Proceedings of the ninth international conference on Information and knowledge management*, pp. 86–93, (2000).

[22] Siyuan Qiao, Wei Shen, Zhishuai Zhang, Bo Wang, and Alan Yuille, 'Deep co-training for semi-supervised image recognition', in *Proceedings of the european conference on computer vision (eccv)*, pp. 135–152, (2018).

[23] Alec Radford, Luke Metz, and Soumith Chintala, 'Unsupervised representation learning with deep convolutional generative adversarial networks', *arXiv preprint arXiv:1511.06434*, (2015).

[24] Kuniaki Saito, Yoshitaka Ushiku, and Tatsuya Harada, 'Asymmetric tri-training for unsupervised domain adaptation', *arXiv preprint arXiv:1702.08400*, (2017).

[25] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen, 'Regularization with stochastic transformations and perturbations for deep semi-supervised learning', *Advances in neural information processing systems*, **29**, (2016).

[26] Reuben R Shamir, Yuval Duchin, Jinyoung Kim, Guillermo Sapiro, and Noam Harel, 'Continuous dice coefficient: a method for evaluating probabilistic segmentations', *arXiv preprint arXiv:1906.11031*, (2019).

[27] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li, 'Fixmatch: Simplifying semi-supervised learning with consistency and confidence', *Advances in neural information processing systems*, **33**, 596–608, (2020).

[28] Antti Tarvainen and Harri Valpola, 'Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results', *Advances in neural information processing systems*, **30**, (2017).

[29] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker, 'Learning to adapt structured output space for semantic segmentation', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7472–7481, (2018).

[30] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell, 'Deep domain confusion: Maximizing for domain invariance', *arXiv preprint arXiv:1412.3474*, (2014).

[31] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros, 'Unpaired image-to-image translation using cycle-consistent adversarial networks', in *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, (2017).

[32] Xiahai Zhuang and Juan Shen, 'Multi-scale patch and multi-modality atlases for whole heart segmentation of mri', *Medical image analysis*, **31**, 77–87, (2016).

[33] Danbing Zou, Qikui Zhu, and Pingkun Yan, 'Unsupervised domain adaptation with dualscheme fusion network for medical image segmentation', in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20, International Joint Conferences on Artificial Intelligence Organization*, pp. 3291–3298, (2020).

[34] Yang Zou, Zhiding Yu, BVK Vijaya Kumar, and Jinsong Wang, 'Unsupervised domain adaptation for semantic segmentation via class-balanced self-training', in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 289–305, (2018).