Unsupervised Cross-Modality Domain Adaptation of ConvNets for Biomedical Image Segmentations with Adversarial Loss

Qi Dou[1*], **Cheng Ouyang**[2*], **Cheng Chen**[1], **Hao Chen**[1,3] and **Pheng-Ann Heng**[1]
[1] Department of Computer Science and Engineering, The Chinese University of Hong Kong
[2] Department of Electrical Engineering and Computer Science, University of Michigan
[3] Imsight Medical Technology Inc., Shenzhen, China
qdou@cse.cuhk.edu.hk, couy@umich.edu, {cchen,hchen,pheng}@cse.cuhk.edu.hk

Abstract

Convolutional networks (ConvNets) have achieved great successes in various challenging vision tasks. However, the performance of ConvNets would degrade when encountering the domain shift. The domain adaptation is more significant while challenging in the field of biomedical image analysis, where cross-modality data have largely different distributions. Given that annotating the medical data is especially expensive, the supervised transfer learning approaches are not quite optimal. In this paper, we propose an unsupervised domain adaptation framework with adversarial learning for cross-modality biomedical image segmentations. Specifically, our model is based on a dilated fully convolutional network for pixel-wise prediction. Moreover, we build a plug-and-play domain adaptation module (DAM) to map the target input to features which are aligned with source domain feature space. A domain critic module (DCM) is set up for discriminating the feature space of both domains. We optimize the DAM and DCM via an adversarial loss without using any target domain label. Our proposed method is validated by adapting a ConvNet trained with MRI images to unpaired CT data for cardiac structures segmentations, and achieved very promising results.

1 Introduction

Deep convolutional networks (ConvNets) have demonstrated great achievements in recent years, achieving state-of-the-art or even human-level performance on various computer vision challenging problems, such as image recognition, semantic segmentation as well as biomedical image diagnosis [He *et al.*, 2016; Esteva *et al.*, 2017]. Typically, the deep networks are trained and tested on datasets where all the samples are drawn from the same probability distribution. However, it has been observed that established models would under-perform when tested on samples from a related but not identical new target domain [Shimodaira, 2000].

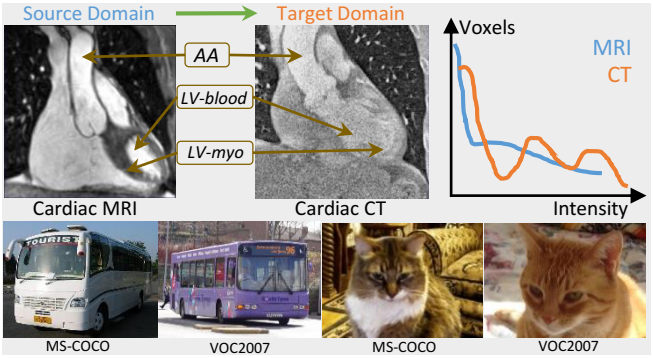

Figure 1: Illustration of severe domain shift existing in cross-modality biomedical images. The appearances of the anatomical structures (AA: ascending aorta, LV-blood: left ventricle blood cavity, LV-myo: left ventricle myocardium) would vary significantly on MRI and CT images. Compared with natural image datasets (see bottom examples), domain adaptation for cross-modality medical images encounter more challenges.

The existence of *domain shift* is common in real-life applications [Gretton *et al.*, 2009; Torralba and Efros, 2011]. The semantic class labels are usually shared between domains, whereas the distributions of data are different. In the field of biomedical image analysis, this issue is even more obvious. Unlike natural images which are generally taken by optical devices, medical radiological images are acquired by different imaging modalities such as Computed Tomography (CT) and Magnetic Resonance Imaging (MRI). Data distributions of these modalities mismatch significantly, due to their different principles of imaging physics. The appearance of anatomical structures are distinct across radiology modalities, with obviously different intensity histograms. In Fig. **??**, we illustrate the severe domain shift between MRI/CT data. In comparison with examples from natural datasets, domain adaptation for cross-modality medical data is more challenging.

To tackle this issue, domain adaptation methods have been studied to generalize the learned models [Patel *et al.*, 2015]. The domain of labeled training data is termed as *source domain*, and the test dataset is called *target domain*. A straightforward solution is transfer learning, i.e., fine-tuning the models learned on source domain with extra labeled data from the target domain [Pan and Yang, 2010]. However, the anno-

[*] Authors contributed equally.

tation is prohibitively time-consuming and expensive, especially for those biomedical datasets. Alternatively, the unsupervised domain adaptation methods are more feasible, given that these scenarios transfer knowledge across domains without using additional target domain labels. Advanced studies in this direction have taken advantage of adversarial training to implicitly learn the feature mapping between domains, and achieved remarkable success in natural datasets [Ganin *et al.*, 2016; Tzeng *et al.*, 2017].

Currently, for biomedical images, how to effectively generalize ConvNets across domains has not yet been fully studied. A representative work is [Kamnitsas *et al.*, 2017] which conducted unsupervised domain adaptation for brain lesion segmentation and achieved promising results. However, their source and target domains are relatively close, given that both are MRI datasets although acquired with different scanners. Adapting ConvNets between cross-modality radiology images with a huge domain shift is more compelling for clinical practice, but has not been explored yet.

In this paper, we propose a novel cross-modality domain adaptation framework for medical image segmentations with unsupervised adversarial learning. To transfer the established ConvNet from source domain (MRI) to target domain (CT) images, we design a plug-and-play domain adaptation module (DAM) which implicitly maps the target input data to the feature space of source domain. Furthermore, we construct a discriminator which is also a ConvNet termed as domain critic module (DCM) to differentiate the feature distributions of two domains. Adversarial loss is derived to train the entire domain adaptation framework in an unsupervised manner, by placing the DAM and DCM into a minimax two-player game. Our main contributions are:

- We pioneer cross-modality domain adaptation for medical image segmentation using deep ConvNets. A flexible plug-and-play framework is designed to transfer a MRI segmenter to CT data via feature-level mapping.
- We optimize our framework with unpaired MRI/CT images via adversarial learning in an unsupervised manner, eliminating the cost of labeling extra medical datasets.
- Extensive experiments with promising results on cardiac segmentation application have validated the feasibility of radiology cross-modality domain adaptation, as well as the effectiveness of our approach towards this task.

2 Related Work

Domain adaptation aims to confront the performance degradation caused by any distribution change occurred after learning a classifier. For deep learning models, this situation also applies, and a trend of studies have been conducted to map the target input to the original source domain or its feature space. In this section, we first present related works of unsupervised domain adaptation that achieved promising results on natural image datasets. Next, we review the recent studies on domain adaptation for medical image segmentations using ConvNets.

Most prior studies on unsupervised domain adaptation focused on aligning the distributions between domains in feature space, by minimizing measures of distance between features extracted from the source and target domains. For example, the Maximum Mean Discrepancy (MMD) was minimized together with a task-specific loss to learn the domain-invariant and semantic-meaningful features in [Tzeng *et al.*, 2014]. The correlations of layer activations between the domains were aligned in the study of [Sun and Saenko, 2016]. Based on this, [Wang *et al.*, 2017] further extended the work and minimized domain difference based on both the first and second order information between source and target domains. Alternatively, with the emergence of generative adversarial network (GAN) [Goodfellow *et al.*, 2014] and its powerful extensions [Radford *et al.*, 2015; Arjovsky *et al.*, 2017], the mapping between domains were implicitly learned via the adversarial loss. The [Ganin *et al.*, 2016] proposed to extract domain-invariant features by sharing weights between two ConvNet classifiers. Later, the [Tzeng *et al.*, 2017] introduced a more flexible adversarial learning method with untied weight sharing, which helps effective learning in the presence of larger domain shifts. Another GAN based direction of solution is to learn a transformation in the pixel space [Bousmalis *et al.*, 2017], adapting the source-domain images to appear as if drawn from the target domain.

In the field of medical image analysis using deep learning, domain adaptation is also an important topic to generalize learned models across data acquired from different imaging protocols. Transfer learning with network fine-tuning strategies has been experimentally studied by [Ghafoorian *et al.*, 2017] on the brain lesion segmentation application. Although the amount was small, annotations from target domain were still required in their scenario. The latest study on medical data that is closely related to our work is [Kamnitsas *et al.*, 2017], which performed unsupervised domain adaptation for brain lesion segmentation. Their ConvNets learned domain-invariant features on images, with an adversarial loss serving as the supervision for feature extraction. The results were inspiring and demonstrated the efficacy of adversarial loss for unsupervised domain adaptation on medical datasets. However, their source and target domains are relatively close, because both were MRI datasets. Although acquired with different scanners and imaging protocols, the images were from the same modality and the domain shift was not dramatic. In contrast, our problem setting, i.e., adapting a ConvNet trained on MRI data to CT images, is novel but more adventurous and challenging, since our domain shift is more severe.

3 Methods

The Fig. 2 presents our proposed framework for unsupervised cross-modality domain adaptation in biomedical image segmentation. Based on a standard ConvNet segmenter, we construct a plug-and-play domain adaptation module (DAM) and a domain critic module (DCM) to form adversarial learning. Details of network architecture, adaptation method, adversarial loss and training strategies are elaborated in this section.

3.1 ConvNet Segmenter Architecture

With the labeled dataset of N^s samples from source domain, denoted by $X^s = \{(x_1^s, y_1^s), ..., (x_{N^s}^s, y_{N^s}^s)\}$, we conduct supervised learning to establish a mapping from the input image to the label space Y^s. In our setting, the x_i^s represents the

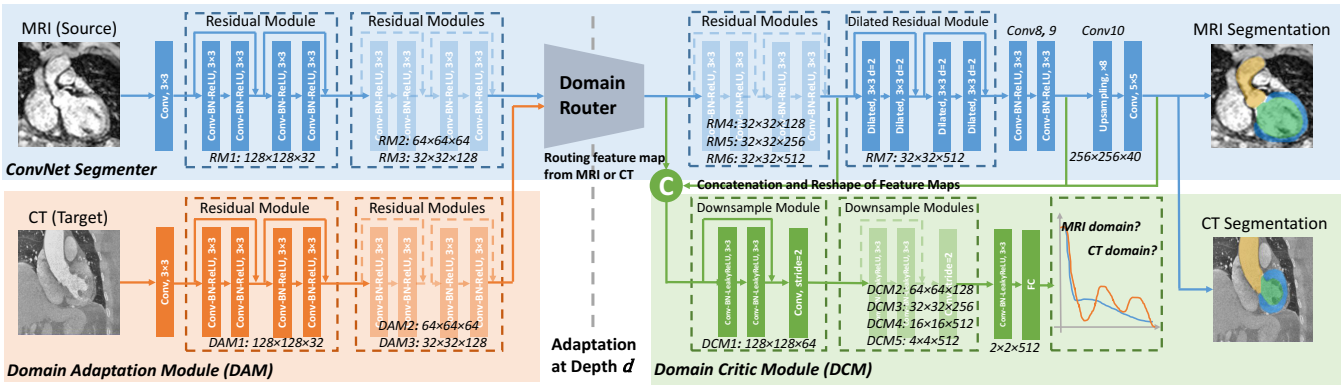

Figure 2: Overview of our proposed plug-and-play framework for cross-modality domain adaptation. The DAM and DCM are optimized via adversarial learning. During inference, the domain router is used for routing feature maps of different domains.

sample (pixel or patch) of medical images and y_i^s is the category of anatomical structures. For the ease of denotation, we omit the index i in the following, and directly use x^s and y^s to represent the samples and labels from the source domain.

The mapping M^s from input to the label space is implicitly learned in the form of a segmentation ConvNet. The backbone of our segmenter is the residual network for pixel-wise prediction of biomedical images. We employ the dilated residual blocks [Yu *et al.*, 2017] to extract representative features from a large receptive field while preserving the spatial acuity of feature maps. More specifically, the image is firstly input to a Conv layer, then forwarded to 3 residual modules (termed as RM, each consisting of 2 stacked residual blocks) and downsampled by a factor of 8. Next, another three RMs and one dilated RM are stacked to form a deep network. To enlarge receptive field for extracting global semantic features, 4 dilated convolutional layers are used in RM7 with a dilation factor of 2. For dense predictions in our segmentation task, we conduct upsamling at layer *Conv10*, which is followed by 5×5 convolutions to smooth out the feature maps. Finally, a softmax layer is used for probability predictions of the pixels.

The segmentation ConvNet using labeled data from source domain is optimized by minimizing the hybrid loss $\mathcal{L}_{\text{seg}}$ composed of the multi-class cross-entropy loss and the Dice coefficient loss [Milletari *et al.*, 2016]. Formally, we denote $y_{i,c}^s$ for binary label regarding class $c \in C$ in sample x_i^s, its probability prediction is $\hat{p}_{i,c}^s$, and the label prediction is $\hat{y}_{i,c}^s$, the source domain segmenter loss function is as follows:

$$\begin{aligned}\mathcal{L}_{\text{seg}} = &-\sum_{i=1}^{N^s}\sum_{c\in C} w_c^s \cdot y_{i,c}^s \log(\hat{p}_{i,c}^s) \\ &-\lambda \sum_{c\in C} \frac{\sum_{i=1}^{N^s} 2y_{i,c}^s \hat{y}_{i,c}^s}{\sum_{i=1}^{N^s} y_{i,c}^s y_{i,c}^s + \sum_{i=1}^{N^s} \hat{y}_{i,c}^s \hat{y}_{i,c}^s},\end{aligned} \tag{1}$$

where the first term is the cross-entropy loss for pixel-wise classification, with w_c^s being a weighting factor to cope with the issue of class imbalance. The second term is the Dice loss for multiple cardiac structures, which is commonly employed in biomedical image segmentation problems. We combine the two complementary loss functions to tackle the challenging heart segmentation task. In practice, we also tried to use only one type of loss, but the performance was not quite high.

3.2 Plug-and-Play Domain Adaptation Module

When the ConvNet is learned on the source domain, our goal is to generalize it to a target domain. In transfer learning, the last several layers of the network are usually fine-tuned for a new task with new label space. The supporting assumption is that early layers in the network extract low-level features (such as edge filters and color blobs) which are common for vision tasks. Those upper layers are more task-specific and learn high-level features for the classifier [Zeiler and Fergus, 2014; Yosinski *et al.*, 2014]. In this case, labeled data from target domain are required to supervise the learning process. Differently, we use unlabeled data from the target domain, given that labeling dataset is time-consuming and expensive. This is critical in clinical practice where radiologists are willing to perform image computing on cross-modality data with as less extra annotation cost as possible. Hence, we propose to adapt the ConvNet with unsupervised learning.

In our segmenter, the source domain mapping M^s is layer-wise feature extractors composing stacked transformations of $\{M_{l_1}^s, ..., M_{l_n}^s\}$, with the l denoting the network layer index. Formally, the predictions of labels are obtained by:

$$\hat{y}^s = M^s(x^s) = M_{l_1:l_n}^s(x^s) = M_{l_n}^s \circ ... \circ M_{l_1}^s(x^s). \tag{2}$$

For domain adaptation, the label space of source and target domains are identical, i.e., we segment the same anatomical structures from medical MRI/CT data. Our hypothesis is that the distribution changes between the cross-modality domains are primarily low-level characteristics (e.g., gray-scale values) rather than high-level (e.g., geometric structures). The higher layers (such as $M_{l_n}^s$) are closely in correlation with the class labels which can be shared across different domains. In this regard, we propose to reuse the feature extractors learned in higher layers of the ConvNet, whereas the earlier layers are updated to conduct distribution mappings in feature space for our unsupervised domain adaptation.

For the input from target domain x^t, we propose a domain adaptation module denoted by $\mathcal{M}$ that maps x^t to the feature space of the source domain. We denote the adaptation depth by d, i.e., the layers earlier than and including l_d are replaced by DAM when processing the target domain images. In the meanwhile, the source model's upper layers are frozen during domain adaptation learning and reused for target inference. Formally, the predictions for target domain is as:

$$\hat{y}^t = M^s_{l_{d+1}:l_n} \circ \mathcal{M}(x^t) = M^s_{l_n} \circ ... \circ M^s_{l_{d+1}} \circ \mathcal{M}(x^t), \quad (3)$$

where $\mathcal{M}(x^t) = \mathcal{M}_{l_1:l_d}(x^t) = \mathcal{M}_{l_d} \circ ... \circ \mathcal{M}_{l_1}(x^t)$ represents the DAM which is also a stacked ConvNet. Overall, we form a flexible plug-and-play domain adaptation framework. During the test inference, the DAM directly replaces the early d layers of the model trained on source domain. The images of target domain are processed and mapped to deep learning feature space of source domain via the DAM. These adapted features are robust to the cross-modality domain shift, and can be mapped to the label space using those high-level layers established on source domain. In practice, the ConvNet configuration of the DAM is identical to $\{M^s_{l_1}, ..., M^s_{l_d}\}$. We initialize the DAM with trained source domain model and fine-tune the parameters in an unsupervised manner with adversarial loss.

3.3 Learning with Adversarial Loss

We propose to train our domain adaptation framework with adversarial loss via unsupervised learning. The spirit of adversarial training roots in GAN, where a generator model and a discriminator model form a minimax two-player game. The generator learns to capture the real data distribution; and the discriminator estimates the probability that a sample comes from the real training data rather than the generated data. These two models are alternatively optimized and compete with each other, until the generator can produce real-like samples that the discriminator fails to differentiate. For our problem, we train the DAM, aiming that the ConvNet can generate source-like feature maps from target input. Hence, the ConvNet is equivalent to a generator from GAN's perspective.

Considering that accurate segmentations come from high-level semantic features, which in turn rely on fine-patterns extracted by early layers, we propose to align multiple levels of feature maps between source and target domains (see Fig. 2). In practice, we select several layers from the frozen higher layers, and refer their corresponding feature maps as the set of $F_H(\cdot)$ where $H = \{k, ..., q\}$ being the set of selected layer indices. Similarly, we denote the selected feature maps of DAM by $\mathcal{M}_A(\cdot)$ with the A being the selected layer set. In this way, the feature space of target domain is $(\mathcal{M}_A(x^t), F_H(x^t))$ and the $(M^s_A(x^s), F_H(x^s))$ is their counterpart for source domain. Given the distribution of $(\mathcal{M}_A(x^t), F_H(x^t)) \sim \mathbb{P}_g$, and that of $(M^s_A(x^s), F_H(x^s)) \sim \mathbb{P}_s$, the distance between these two domain distributions which needs to be minimized is represented as $W(\mathbb{P}_s, \mathbb{P}_g)$. For stabilized training, we employ the Wassertein distance [Arjovsky *et al.*, 2017] between the two distributions as follows:

$$W(\mathbb{P}_s, \mathbb{P}_g) = \inf_{\gamma \sim \prod(\mathbb{P}_s, \mathbb{P}_g)} \mathbb{E}_{(\mathrm{x},\mathrm{y}) \sim \gamma}[\|\mathrm{x} - \mathrm{y}\|], \quad (4)$$

where $\prod(\mathbb{P}_s, \mathbb{P}_g)$ represents the set of all joint distributions $\gamma(\mathrm{x}, \mathrm{y})$ whose marginals are respectively $\mathbb{P}_s$ and $\mathbb{P}_g$.

In adversarial learning, the DAM is pitted against an adversary: a discriminative model that implicitly estimates the $W(\mathbb{P}_s, \mathbb{P}_g)$. We refer our discriminator as domain critic module and denote it by $\mathcal{D}$. Specifically, our constructed DCM consists of several stacked residual blocks, as illustrated in Fig. 2. In each block, the number of feature maps is doubled until it reaches 512, while their sizes are decreased. We concatenate the multiple levels of feature maps as input to the DCM. This discriminator would differentiate the complicated feature space between the source and target domains. In this way, our domain adaptation approach not only removes source-specific patterns in the beginning but also disallows their recovery at higher layers [Kamnitsas *et al.*, 2017]. In unsupervised learning, we jointly optimize the generator $\mathcal{M}$ (DAM) and the discriminator $\mathcal{D}$ (DCM) via adversarial loss. Specifically, with X^t being target set, the loss for learning the DAM is:

$$\min_{\mathcal{M}} \mathcal{L}_{\mathcal{M}}(X^t, \mathcal{D}) = -\mathbb{E}_{(\mathcal{M}_A(x^t), F_H(x^t)) \sim \mathbb{P}_g}[\mathcal{D}(\mathcal{M}_A(x^t), F_H(x^t))]. \quad (5)$$

Furthermore, with the X^s representing the set of source images, the DCM is optimized via:

$$\min_{\mathcal{D}} \mathcal{L}_{\mathcal{D}}(X^s, X^t, \mathcal{M}) = \mathbb{E}_{(\mathcal{M}_A(x^t), F_H(x^t)) \sim \mathbb{P}_g}[\mathcal{D}(\mathcal{M}_A(x^t), F_H(x^t))] - \mathbb{E}_{(M^s_A(x^s), F_H(x^s)) \sim \mathbb{P}_s}[\mathcal{D}(M^s_A(x^s), F_H(x^s))], s.t.\ \|\mathcal{D}\|_{L \leq K}, \quad (6)$$

where K is a constant that applies Lipschitz contraint to $\mathcal{D}$.

During the alternative updating of $\mathcal{M}$ and $\mathcal{D}$, the DCM outputs a more precise estimation of $W(\mathbb{P}_s, \mathbb{P}_g)$ between distributions of the feature space from both domains. The updated DAM is more effective to generate source-like feature maps for conducting cross-modality domain adaptation.

3.4 Training Strategies

In our setting, the source domain is biomedical cardiac MRI images and the target domain is CT data. All the volumetric MRI and CT images were re-sampled to the voxel spacing of $1 \times 1 \times 1$ mm^3 and cropped into the size of $256 \times 256 \times 256$ centering at the heart region. In preprocessing, we conducted intensity standardization for each domain, respectively. Augmentations of rotation, zooming and affine transformations were employed to combat over-fitting. To leverage the spatial information existing in volumetric data, we sampled consecutive three slices along the coronal plane and input them to three channels. The label of the intermediate slice is utilized as the ground truth when training the 2D networks.

We first trained the segmenter on the source domain data in supervised manner with stochastic gradient descent. The Adam optimizer was employed with parameters as batch size of 5, learning rate of 1×10^{-3} and a stepped decay rate of 0.95 every 1500 iterations. After that, we alternatively optimized the DAM and DCM with the adversarial loss for unsupervised domain adaptation. Following the heuristic rules of training WGAN [Arjovsky *et al.*, 2017], we updated the DAM every 20 times when updating the DCM. In adversarial learning, we utilized the RMSProp optimizer with a learning rate of 3×10^{-4} and a stepped decay rate of 0.98 every 100 joint updates, with weight clipping for the discriminator being 0.03.

4 Experiment

4.1 Dataset and Evaluation Metrics

We validated our proposed unsupervised cross-modality domain adaptation method for biomedical image segmentations on the public dataset of *MICCAI 2017 Multi-Modality Whole*

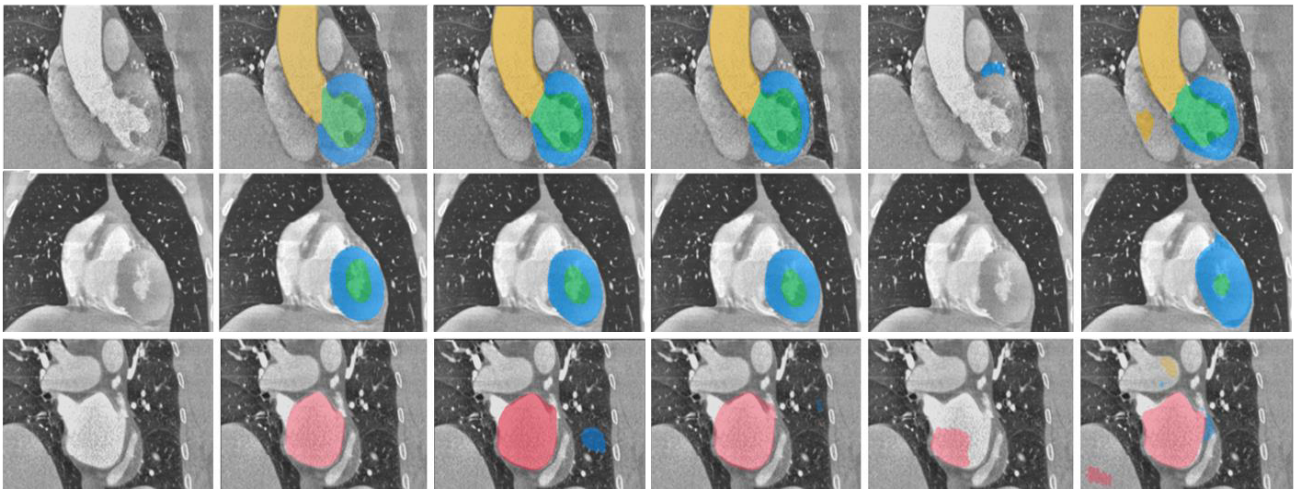

Figure 3: Results of different methods for CT image segmentations. Each row presents one typical example, from left to right: (a) raw CT slices (b) ground truth labels (c) supervised transfer learning (d) ConvNets trained from scratch (e) directly applying MRI segmenter on CT data (f) our unsupervised cross-modality domain adaptation results. The structures of AA, LA-blood, LV-blood and LV-myo are indicated by yellow, red, green and blue colors, respectively (best viewed in color).

Heart Segmentation [Zhuang and Shen, 2016]. This dataset consists of unpaired 20 MRI and 20 CT images from 40 patients. The MRI and CT data were acquired in different clinical centers. The cardiac structures of the images were manually annotated by radiologists for both MRI and CT images. Our ConvNet segmenter aimed to automatically segment four cardiac structures including the ascending aorta (AA), the left atrium blood cavity (LA-blood), the left ventricle blood cavity (LV-blood), and the myocardium of the left ventricle (LV-myo). For each modality, we randomly split the dataset into training (16 subjects) and testing (4 subjects) sets, which were fixed throughout all experiments.

For evaluation metrics, we followed the common practice to quantitatively evaluate the segmentation performance for automatic methods [Dou *et al.*, 2017]. The DICE coefficient ($[\%]$) was employed to assess the agreement between the predicted segmentation and ground truth for cardiac structures. We also calculated the average surface distance (ASD$[$voxel$]$) to measure the segmentation performance from the perspective of the boundary. A higher Dice and lower ASD indicate better segmentation performance. Both metrics are presented in the format of *mean*$\pm$*std*, which shows the average performance as well as the cross-subject variations of the results.

4.2 Experimental Settings

In our experiments, the source domain is the MRI images and the target domain is the CT dataset. We demonstrated the effectiveness of the proposed unsupervised cross-modality domain adaptation method with extensive experiments. We designed several experiment settings: 1) training and testing the ConvNet segmenter on source domain (referred as *Seg-MRI*); 2) training the segmenter from scratch on annotated target domain data (referred as *Seg-CT*); 3) fine-tuning the source domain segmenter with annotated target domain data, i.e., the supervised transfer learning (referred as *Seg-CT-STL*); 4) directly testing the source domain segmenter on target domain data (referred as *Seg-CT-noDA*); 5) our proposed unsupervised domain adaptation method (referred as *Seg-CT-UDA*). We also compared with a previous state-of-the-art heart segmentation method using ConvNets [Payer *et al.*, 2017]. Last but not least, we conducted ablation studies to observe how the adaptation depth would affect the performance.

4.3 Results of Unsupervised Domain Adaptation

The results of different methods are listed in Table 1, which demonstrates that the proposed unsupervised domain adaptation method is effective by mapping the feature space of target CT domain to that of source MRI domain. Qualitative results of the segmentations for CT images are presented in Fig. 3.

We first evaluate the performance of the segmenter for *Seg-MRI*, which is the source domain model and serves as the basis for subsequent domain adaptation procedures. Compared with the [Payer *et al.*, 2017], our ConvNet segmenter reached promising performance with exceeding Dice on LV-blood and LV-myo, as well as comparable Dice on AA and LA-blood. With this standard segmenter network architecture, we conducted following experiments to validate the effectiveness of our unsupervised domain adaptation framework.

To experimentally explore the potential upper-bounds of the segmentation accuracy of the cardiac structures from CT data, we implemented two different settings, i.e., the *Seg-CT* and *Seg-CT-STL*. Generally, the segmenter fine-tuned from *Seg-MRI* achieved higher Dice and lower ASD than the model trained from scratch, proving the effectiveness of supervised transfer learning for adapting an established network to a related target domain using additional annotations. Meanwhile, these results are comparable to [Payer *et al.*, 2017] on most of the four cardiac structures.

As for observing the severe domain shift problem inherent in cross-modality biomedical images, we directly applied the segmenter trained on MRI domain to the CT data without any domain adaptation procedure. Unsurprisingly, the network of *Seg-MRI* completely failed on CT images, with average Dice of merely 14.3% across the structures. As shown in Table 1,

Methods	AA		LA-blood		LV-blood		LV-myo	
	Dice	ASD	Dice	ASD	Dice	ASD	Dice	ASD
DL-MR [Payer *et al.*, 2017]	76.6±13.8	-	81.1±13.8	-	87.7±7.7	-	75.2±12.1	-
DL-CT [Payer *et al.*, 2017]	91.1±18.4	-	92.4±3.6	-	92.4±3.3	-	87.2±3.9	-
Seg-MRI	75.9±5.5	12.9±8.4	78.8±6.8	16.0±8.1	90.3±1.3	2.0±0.2	75.5±3.6	2.6±1.4
Seg-CT	81.3±24.4	2.1±1.1	89.1±3.0	10.6±6.9	88.8±3.7	21.3±8.8	73.3±5.9	42.8±16.4
Seg-CT-STL	78.3±2.8	2.9±2.0	89.7±3.6	7.6±6.7	91.6±2.2	4.9±3.2	85.2±3.3	5.9±3.8
Seg-CT-noDA	19.7±2.0	31.2±17.5	25.7±17.2	8.7±3.3	0.8±1.3	N/A	11.1±14.4	31.0±37.6
Seg-CT-UDA (d=13)	63.9±15.4	**13.9±5.6**	54.7±13.2	16.6±6.8	35.1±26.1	**18.4±5.1**	35.4±18.4	**14.2±5.3**
Seg-CT-UDA (d=21)	**74.8±6.2**	27.5±7.6	51.1±11.2	20.1±4.5	**57.2±12.4**	29.5±11.7	**47.8±5.8**	31.2±10.1
Seg-CT-UDA (d=31)	71.9±0.5	25.8±12.5	**55.2±22.9**	**15.2±8.2**	39.2±21.8	21.2±3.9	34.3±19.1	24.7±10.5

Table 1: Quantitative comparison of segmentation performance on cardiac structures between different methods. (Note: the - means that the results were not reported by that method.)

the *Seg-CT-noDA* only got a Dice of 0.8% for the LV-blood. The model did not even output any correct predictions for two of the four testing subjects on the structure of LV-blood (please refer to (e) in Fig. 3). This demonstrates that although the cardiac MRI and CT images share similar high-level representations and identical label space, the significant difference in their low-level characteristics makes it extremely difficult for MRI segmenter to extract effective features for CT.

With our unsupervised domain adaptation method, we find a great improvement of the segmentation performance on the target CT data compared with the *Seg-CT-noDA*. More specifically, our *Seg-CT-UDA (d=21)* model has increased the average Dice across four cardiac structures by 43.4%. As presented in Fig. 3, the predicted segmentation masks from *Seg-CT-UDA* can successfully localize the cardiac structures and further capture their anatomical shapes. The performance on segmenting AA is even close to that of *Seg-CT-STL*. This reflects that the distinct geometric pattern and the clear boundary of the AA have been successfully captured by the DCM. In turn, it supervises the DAM to generate similar activation patterns as the source feature space via adversarial learning. Looking at the other three cardiac structures (i.e., LA-blood, LV-blood and LV-myo), the *Seg-CT-UDA* performances are not as high as that of AA. The reason is that these anatomical structures are more challenging, given that they come with either relatively irregular geometrics or limited intensity contrast with surrounding tissues. The deficiency focused on the unclear boundaries between neighboring structures or noise predictions on relatively homogeneous tissues away from the ROI. This is responsible for the high ASDs of *Seg-CT-UDA*, where boundaries are corrupted by noisy outputs. Nevertheless, by mapping the feature space of target domain to that of the source domain, we obtained greatly improved and promising segmentations against *Seg-CT-noDA* with zero data annotation effort.

4.4 Ablation Study on Adaptation Depth

The adaptation depth d is an important hyper-parameter in our framework, which determines how many layers to be replaced during the plug-and-play domain adaptation procedure. Intuitively, a shallower DAM (i.e., smaller d) might be less capable of learning effective feature mapping function $\mathcal{M}$ across domains than a deeper DAM (i.e., larger d). This is due to the insufficient capacity of parameters in shallow DAM, as well as the huge domain shift in feature distributions. Conversely, with an increase in adaptation depth d, DAM becomes more powerful for feature mappings, but training a deeper DAM solely with adversarial gradients would be more challenging. Towards this issue, we conducted ablation studies to demonstrate how the performance would be affected by d.

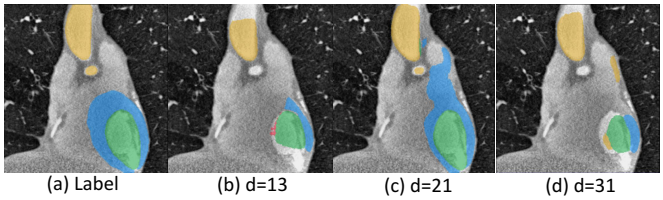

Figure 4: Comparison of results using *Seg-CT-UDA* with different adaptation depth (colors are the same with Fig. 3).

To validate above intuitions and search for an optimal d, we repeated the experiment with domain adaptation from MRI to CT by varying the $d = \{13, 21, 31\}$, while maintaining all the other settings the same. Viewing the examples in Fig. 4, *Seg-CT-UDA (d=21)* model obtained an approaching ground-truth segmentation mask for ascending aorta. The other two models also produced inspiring results capturing the geometry and boundary characteristics of *AA*, validating the effectiveness of our unsupervised domain adaptation method. From the Table 1, we can observe that DAM with a middle-level of adaptation depth (*d=21*) achieved the highest Dice on three of the four cardiac structures, exceeding the other two models by a significant margin. For the LA-blood, the three adaptation depths reached comparable segmentation Dice and ASD, and the *d=31* model was the best. Notably, the model of *Seg-CT-UDA (d=31)* overall demonstrated superiority over the model with adaptation depth *d=13*. This shows that enabling more layers learnable helps to improve the domain adaptation performance on cross-modality segmentations.

5 Conclusion

This paper pioneers to propose an unsupervised domain adaptation framework for generalizing ConvNets across different modalities of biomedical images. The flexible plug-and-play framework is obtained by optimizing a DAM and DCM via adversarial learning. Extensive experiments with promising results on cardiac segmentations have validated the effectiveness of our approach.

Acknowledgments

The work described in this paper was supported by the following grants from Hong Kong Research Grants Council under General Research Fund Scheme (Project no. 14202514 and 14203115).

References

[Arjovsky *et al.*, 2017] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

[Bousmalis *et al.*, 2017] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *CVPR*, 2017.

[Dou *et al.*, 2017] Qi Dou, Lequan Yu, Hao Chen, Yueming Jin, Xin Yang, Jing Qin, and Pheng-Ann Heng. 3d deeply supervised network for automated segmentation of volumetric medical images. *Medical image analysis*, 41:40–54, 2017.

[Esteva *et al.*, 2017] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017.

[Ganin *et al.*, 2016] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

[Ghafoorian *et al.*, 2017] Mohsen Ghafoorian, Alireza Mehrtash, Tina Kapur, Nico Karssemeijer, Elena Marchiori, Mehran Pesteie, Charles RG Guttmann, Frank-Erik de Leeuw, Clare M Tempany, Bram van Ginneken, et al. Transfer learning for domain adaptation in mri: Application in brain lesion segmentation. In *MICCAI*, pages 516–524, 2017.

[Goodfellow *et al.*, 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014.

[Gretton *et al.*, 2009] Arthur Gretton, Alexander J Smola, Jiayuan Huang, Marcel Schmittfull, Karsten M Borgwardt, and Bernhard Schölkopf. Covariate shift by kernel mean matching. 2009.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[Kamnitsas *et al.*, 2017] Konstantinos Kamnitsas, Christian Baumgartner, Christian Ledig, Virginia Newcombe, Joanna Simpson, Andrew Kane, David Menon, Aditya Nori, Antonio Criminisi, Daniel Rueckert, et al. Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. In *International Conference on Information Processing in Medical Imaging*, pages 597–609. Springer, 2017.

[Milletari *et al.*, 2016] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *3D Vision (3DV), 2016 Fourth International Conference on*, pages 565–571. IEEE, 2016.

[Pan and Yang, 2010] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010.

[Patel *et al.*, 2015] Vishal M Patel, Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. Visual domain adaptation: A survey of recent advances. *IEEE signal processing magazine*, 32(3):53–69, 2015.

[Payer *et al.*, 2017] Christian Payer, Darko Štern, Horst Bischof, and Martin Urschler. Multi-label whole heart segmentation using cnns and anatomical label configurations. pages 190–198, 2017.

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

[Shimodaira, 2000] Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

[Sun and Saenko, 2016] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *Proceedings of the ECCV Workshops*, pages 443–450. Springer, 2016.

[Torralba and Efros, 2011] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR*, pages 1521–1528, 2011.

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

[Tzeng *et al.*, 2017] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *CVPR*, pages 2962–2971, 2017.

[Wang *et al.*, 2017] Yifei Wang, Wen Li, Dengxin Dai, and Luc Van Gool. Deep domain adaptation by geodesic distance minimization. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017.

[Yosinski *et al.*, 2014] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *NIPS*, pages 3320–3328, 2014.

[Yu *et al.*, 2017] Fisher Yu, Vladlen Koltun, and Thomas Funkhouser. Dilated residual networks. In *CVPR*, pages 636–644, 2017.

[Zeiler and Fergus, 2014] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *ECCV*, pages 818–833. Springer, 2014.

[Zhuang and Shen, 2016] 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.