

Structure-Driven Unsupervised Domain Adaptation for Cross-Modality Cardiac Segmentation

Zhiming Cui^{ID}, Changjian Li^{ID}, Zhixu Du^{ID}, Nenglun Chen, Guodong Wei^{ID}, Runnan Chen, Lei Yang^{ID}, Dinggang Shen^{ID}, *Fellow, IEEE*, and Wenping Wang, *Fellow, IEEE*

Abstract—Performance degradation due to domain shift remains a major challenge in medical image analysis. Unsupervised domain adaptation that transfers knowledge learned from the source domain with ground truth labels to the target domain without any annotation is the mainstream solution to resolve this issue. In this paper, we present a novel unsupervised domain adaptation framework for cross-modality cardiac segmentation, by explicitly capturing a common cardiac structure embedded across different modalities to guide cardiac segmentation. In particular, we first extract a set of 3D landmarks, in a self-supervised manner, to represent the cardiac structure of different modalities. The high-level structure information is then combined with another complementary feature, the Canny edges, to produce accurate cardiac segmentation results both in the source and target domains. We extensively evaluate our method on the MICCAI 2017 MM-WHS dataset for cardiac segmentation. The evaluation, comparison and comprehensive ablation studies demonstrate that our approach achieves satisfactory segmentation results and outperforms state-of-the-art unsupervised domain adaptation methods by a significant margin.

Index Terms—Cross-modality learning, unsupervised domain adaptation, structure distillation, cardiac segmentation.

Manuscript received March 31, 2021; revised May 14, 2021; accepted May 19, 2021. Date of publication June 23, 2021; date of current version November 30, 2021. (Corresponding authors: Changjian Li; Dinggang Shen; Wenping Wang.)

Zhiming Cui is with the School of Biomedical Engineering, ShanghaiTech University, Shanghai 201210, China, and also with the Department of Computer Science, The University of Hong Kong, Hong Kong (e-mail: zmcui@cs.hku.hk).

Changjian Li is with the Department of Computer Science, University College London, London WC1E 6EA, U.K. (e-mail: chjili2011@gmail.com).

Zhixu Du, Nenglun Chen, Guodong Wei, Runnan Chen, Lei Yang, and Wenping Wang are with the Department of Computer Science, The University of Hong Kong, Hong Kong (e-mail: dzx3501@connect.hku.hk; chennenglun@gmail.com; g.d.wei.china@gmail.com; runnanchen@modontics.com; yanglei.dalian@gmail.com; wenping@cs.hku.hk).

Dinggang Shen is with the School of Biomedical Engineering, ShanghaiTech University, Shanghai 201210, China, also with Shanghai United Imaging Intelligence Company Ltd., Shanghai 200030, China, and also with the Department of Artificial Intelligence, Korea University, Seoul 02841, South Korea (e-mail: dinggang.shen@gmail.com).

Digital Object Identifier 10.1109/TMI.2021.3090432

I. INTRODUCTION

WITH the recent development of deep learning techniques, great success has been achieved on various challenging tasks in medical image analysis, such as segmentation, detection and diagnosis [1]–[3], reaching even human-level performances when testing samples are collected following the same distribution as training data. However, generalizing a well-trained model to new domains is difficult due to the domain shift. It is a common issue especially when applying the trained model to real-world clinical scenarios, since medical images are usually captured with different physical properties. For example, Magnetic Resonance Imaging (MRI) and Computed Tomography (CT) images of cardiac play complementary roles in clinical diagnosis, while they have significantly different appearances as shown in Fig. 1. Considering the tedious and expensive annotation by experienced experts, unsupervised domain adaptation that transfers knowledge from a source domain to a target one is an appealing yet challenging solution to the problem.

Previous work tackle this problem by using Generative Adversarial Networks (GANs) [4] to align either the image appearances (e.g., [5], [6]) or their latent features (e.g., [7], [8]) to achieve unsupervised domain adaptation in medical image analysis. The former manipulates the image appearance in the target domain to have a similar style with those from the source domain, while the latter learns to extract the domain invariant latent features. We argue that since medical images of different modalities have significantly different appearances (see CT and MRI as shown in Fig. 1), simply manipulating the styles or latent features of images may not guarantee good domain adaptation results. Furthermore, however, medical images of different modalities are captured to reveal the *same* anatomical structures, thus *explicitly* constraining the neural models with respect to such common anatomical structures across images of different modalities may benefit the domain adaptation. While the definition of such anatomical structures vary across tasks and is not available for most of medical image datasets, this observation motivates the following design.

We propose a deep neural model that learns and leverages the consistent structure in human-body anatomy as high-level

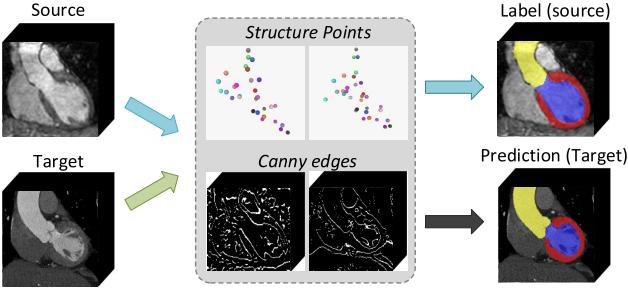


Fig. 1. Overview of our unsupervised domain adaptation method on cardiac MRI and CT images. The example shows MRI (source) to CT (target) adaptation. Specifically, our method aims to extract the cardiac structure across modalities by a set of landmarks, while accurately predict object labels in the target domain guided by the domain-invariant structural landmarks and edge information.

guidance for downstream tasks. Specifically, inspired by the unsupervised discovery from faces [9] and human bodies [10] representing the object geometry without manual annotation, we describe the consistent structure embedded in different modality images as a set of 3D keypoints. This formulation requires no supervision from human-annotated landmark coordinates, and the network can discover consistent 3D keypoints across images of different modalities in a self-supervised manner.

In this paper, we showcase this novel structure-driven approach to domain adaptation for unsupervised cross-modality cardiac segmentation. The proposed framework consists of two core modules, the unsupervised 3D landmark discovery module and the cardiac segmentation module. The first module is designed to capture the explicit structural representation shared across modalities, while the second is designed to obtain the accurate object labels guided by the domain-invariant anatomical structure. Specifically, to extract a consistent structure represented by 3D landmarks, we resort to using a conditional image generation mechanism that aims to reconstruct the target image by the landmarks extracted from itself and the appearance from its deformed version. The deformed image is randomly distorted from the source image so as to have the same appearance but a deformed structure. In order to delineate accurate organ boundaries, we further incorporate in the segmentation module the low-level Canny edges [11] as the complementary features in addition to the domain invariant 3D landmarks that serves as high-level structure information of the content. Furthermore, to preserve the domain consistency in landmark generation and edge extraction steps, we exploit adversarial training in the end-to-end learning framework.

Our main contributions are summarized as follows:

- We present a novel unsupervised domain adaptation approach based on the learned domain invariant structure which plays an essential role to address the severe domain shift problem.
- We distill the 3D landmarks from source and target domains by unsupervised conditional image generation,

and incorporate Canny edges as hybrid information to generate accurate segmentation results.

- Extensive experiments on the dataset of MM-WHS challenge [12] demonstrate that our approach achieves state-of-the-art performance of bidirectional domain adaptation between MR and CT images.

II. RELATED WORK

A. Unsupervised Domain Adaption

Domain adaptation is an important research field to solve performance degradation caused by inter-scanner or cross-modality variations in medical image analysis [13]–[16]. Many deep learning-based methods [7], [17]–[19] have been proposed to transfer the knowledge learned from the source domain to the target domain in a supervised or unsupervised way. However, unsupervised domain adaptation without target domain labels is more desirable and related to our work, we focus on this category in the following.

With the current advances of deep learning techniques, plenty works adopt adversarial learning [4], [5] to address the domain shift problem. They can be mainly divided into three categories: feature-level alignment [7], [20]–[25], image-level alignment [5], [6], [26]–[29] and their mixtures [19], [30]–[32]. Feature-level alignment aims to extract domain-invariant features from source and target images. For example, Tzeng *et al.* [8] apply a discriminator to distinguish the features across modalities via adversarial learning. Kamnitsas *et al.* [33] propose a multi-connected domain discriminator to improve the feature alignment in brain lesion segmentation. Similar works [7], [20]–[25] align the cross-modality features extracted from either semantic segmentation space or image space, demonstrating more effective feature-level alignment.

Following the success of CycleGAN [5] on image-to-image translation tasks, [6], [26]–[29] align image appearance across modalities by translating the image style from the source domain to target domain. For instance, Jiang *et al.* [28] present a tumor-aware unsupervised domain adaptation network to transform CT images to MRI images for lung cancer segmentation. Although alignments on feature or image level achieve promising results in unsupervised domain adaptation, the combination of these two techniques exhibits favorable performance [19], [30]–[32]. In particular, Chen *et al.* [19], [30] conduct synergistic alignment of domains to perform bidirectional unsupervised domain adaptation between cardiac CT and MRI images, which aligns both image appearance and latent features and achieves leading performance.

Although several style or latent feature adaptation methods have already achieved good performance, there is still a large performance gap for such kinds of methods reaching a promising result in the specific cardiac segmentation task. We argue that the extracted domain-invariant style or latent features of images are implicit, and it is hard to tell exactly what the network has learned. Instead, medical images of different modalities are scanned to reveal the same anatomical structures, explicitly detecting the common anatomical

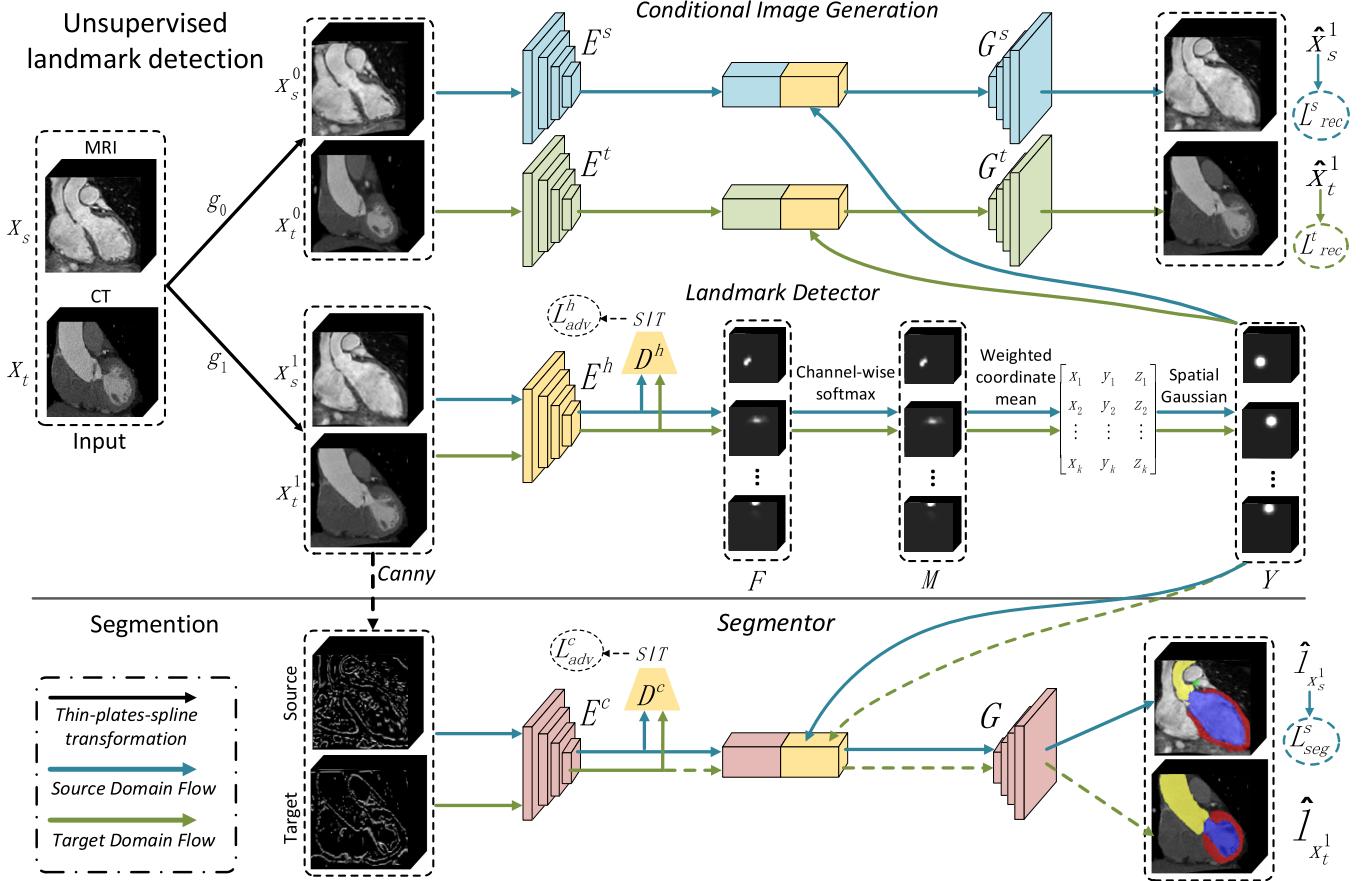


Fig. 2. The framework of our method for unsupervised domain adaptation. The landmark detection module first extracts the anatomical cardiac structure represented by a set of 3D landmarks. The disentangle process is achieved by the conditional image generation mechanism. Then, the segmentation module takes as input both the extracted structure and a set of edges from the canny operator to produce reliable results.

structure with clear meaning, e.g., the left corner point of the left atrium blood cavity, is more practical in the cardiac segmentation task of different modalities.

Recently, domain-invariant information disentanglement has attracted many research attention for unsupervised domain adaptation [34]–[40]. These methods distill most informative and decisive factors across domains, which is less affected by domain specific information in semantic segmentation, e.g., image textures. For example, Chang *et al.* [34] learn structural contents for image translation from synthetic to real-world driving scenes. We share a similar idea with these methods but have significant differences, i.e., in medical images, the anatomical structures of human-bodies are consistent in different image modalities. Thus, we explicitly use a set of spatial points, instead of latent features or image appearances, to represent the organ geometric structures across domains, which better preserves the domain-invariant property.

B. Cross-Modality Cardiac Segmentation

Cardiac MRI and CT images play complementary roles for heart disease diagnosis [41] in real-world clinics. MRI is a gold standard to evaluate the cardiac function, where CT images is an effective indicator for atherosclerosis or coronary artery disease. Recently, with the development of deep learning techniques, more attention has been drawn

to cross-modality learning of cardiac MRI and CT images. Dou *et al.* [7] and Chen *et al.* [19] propose a stream of unsupervised cardiac segmentation networks based on feature- or image-level alignment. Bian *et al.* [42] boost the segmentation performance using an uncertain-aware attention module emphasizing cardiac boundaries. In addition, Li *et al.* [43] present the mutual knowledge distillation scheme to exploit the shared knowledge between cardiac MRI and CT images, while Dou *et al.* [44] share the similar idea by improving network parameter efficiency and proposing a new knowledge distillation loss term.

III. METHOD

Fig. 2 shows an overview of our unsupervised domain adaptation framework for cardiac segmentation. It is composed of two main modules, i.e., the unsupervised landmark detection module and the structure-driven segmentation module. The former aims to discover a set of 3D landmarks that reflect the shared structure across different modalities of cardiac images while the later is designed to obtain the final cardiac segmentation results.

A. Unsupervised Landmark Detection

While most of previous works rely on the implicit alignment to achieve domain adaptation, we propose to leverage an

ordered set of explicitly learned 3D landmarks that are shared across modalities to transfer the knowledge across domains. Such points are learned in a self-supervised manner without human definition, but they can represent the cardiac structure with consistent anatomical locations among different images, as well as across modalities.

We formulate this problem as conditional image generation, following the previous works [10], [45] on facial landmark localization, by treating the learned structural landmarks as the generation condition, or high-level guidance. Specifically, we build an image pair (x^0, x^1) by applying two random geometric deformations to image x . This derives a pair of images sharing the same visual appearance but with deformed geometric structures. Then a decoder is employed to reconstruct image x^1 , so that the generated image x^1 shares the same appearance style with x_0 (by taking x_0 as style input) while retaining the structure of x_1 (by taking its landmarks). This enables the network to learn the 3D structural landmarks by disentangling the structure from the appearance in an self-supervised manner. In the following, we elaborate our approach to achieve the proposed design.

1) Image Pairs Building: We first preprocess input images from the source or target domains to build a pair of deformed images. Formally, the input images X consist of labelled data $\{X_s\}$ from the source domain with corresponding labels $\{L_s\}$, and unlabelled data $\{X_t\}$ from the target domain. For each image $x \in X$, we generate a data pair $(x^0, x^1) = (g_0(x), g_1(x))$ by applying two random 3D thin-plate-spline (TPS) geometric deformations g_0 and g_1 as shown in Fig. 2. Thus, the paired images x^0 and x^1 hold the same appearance style but different geometric structures [10], [46], [47]. The image pairs are then used as input for the unsupervised landmark detection module to extract 3D landmarks consistently across the source and target domains.

2) 3D Landmark Detector: The 3D landmark detector takes as input an image x^1 from the source or target domain and outputs K consistent landmarks to represent the underlying structure. In particular, a standard convolutional neural network, E^h , is employed to encode the input image x^1 (a source or target domain image) into a K -channel feature map $F^{H \times W \times D \times K}$, where each channel $F(k)$ ($k = 1, \dots, K$) is then converted into a probability map $M(k)^{H \times W \times D}$ by channel-wise softmax normalization. To generate the 3D landmarks and ensure the step is differentiable for back-propagation, we use the soft-argmax operator [48] on the k -th probability map $M(k)$ to generate a weighted sum coordinate $c^*(k) \in \mathbb{R}^3$ as the k -th landmark, defined as:

$$M_c(k) = \frac{\exp(F_c(k))}{\sum_{c \in \Omega_k} \exp(F_c(k))}, \quad (1)$$

$$c^*(k) = \sum_{c \in \Omega_k} c \cdot M_c(k), \quad (2)$$

where $c \in \Omega_k$ denotes the spatial coordinates, and $F_c(k)$ is the feature value of the coordinate c referencing to the k -th feature map $F(k)$, while $M_c(k)$ is the weight (i.e., the probability value) of coordinate c over all the other coordinates in the k -th probability map $M(k)$. We then convert the 3D coordinates of the detected landmarks $\{c^*\}$ to K Gaussian maps $Y(k)^{H \times W \times D}$

centered at $c^*(k)$ with a small fixed standard deviation σ , defined as:

$$Y_c(k) = \exp\left(-\frac{1}{2\sigma^2} \|c - c^*(k)\|^2\right), \quad (3)$$

where the standard deviation σ is set to 0.8 for all the experiments. The rationale behind this design is two-fold: on the one hand, using a Gaussian map image to represent a landmark is more conducive than a vector of real numbers (the landmark coordinates), in the meanwhile it is easy to be plugged into the internal features without further Fc layers; on the other, using the probability maps $M(k)$ directly may introduce appearance information to the landmark detection process, which we desire to factor out. Thus, representing the learned 3D landmarks by Gaussian maps centered at the landmark coordinate shall retain the structural information learned from image x^1 with a dense signal while minimizing the effect of image appearance, leading to a disentangled representation. Note here, each map in F , M and Y , has the same size of $H \times W \times D$, which is smaller than the input and output image size, due to the existence of the encoder.

As the networks accept input from both domains, the parameters of the 3D landmark detector are shared to ensure the same landmark semantics across modalities (i.e. CT and MRI). Additionally, to facilitate the same purpose we adopt the adversarial learning strategy to align the feature maps output by the encoder. Adversarial learning is achieved by introducing an auxiliary discriminator D^h , as shown in Fig. 2 and is defined as:

$$\begin{aligned} \mathcal{L}_{adv}^h &= \mathbb{E}_{X_s} \left[\log \left(D^h \left(E^h(x_s^1) \right) \right) \right] \\ &\quad + \mathbb{E}_{X_t} \left[\log \left(1 - D^h \left(E^h(x_t^1) \right) \right) \right], \end{aligned} \quad (4)$$

where \mathbb{E}_{X_s} refers to the expected value over all the source domain images x_s , and \mathbb{E}_{X_t} is the expected value over target domain images x_t .

3) Conditional Image Generation: To achieve the self-supervised landmark learning, we adopt the conditional image generation network (see Fig. 2) to reproduce image x^1 with inputs of image x^0 and the structural landmarks Y distilled from x^1 . Since the paired images (x^0, x^1) are generated from the same image x with different geometric transformations, using either the image x^0 or the landmark-based Gaussian map Y should not be able to reproduce image x^1 . To this end, the network has to utilize the visual appearance feature of x^0 and combine it with the distilled structural information of x^1 to reconstruct image x^1 faithfully. Considering the different visual appearance embedded in the source and target domain images, as shown in Fig. 2, different encoders E^s , E^t and decoders D^s , D^t are employed in the conditional image generation process.

To supervise the landmark detection network, the voxel-based mean squared error (MSE) between the generated image and its ground truth image is minimized. It is worthy noting that, in our segmentation scenario, to make the structural landmark information only relevant to the cardiac region, we discard the reconstruction error of the background region. Specifically, for images from the source domain, we compute

only the MSE loss over the foreground regions which are masked by the ground truth labels of the cardiac regions, denoted as \mathcal{L}_{rec}^s . While for images from the target domain, since the ground truth masks are not available, we instead supervise the reconstruction error over the predicted cardiac regions produced by the segmentor (introduced later in Sec. III-B), denoted as \mathcal{L}_{rec}^t .

B. Structure-Driven Segmentation

With the learned structural landmarks, we further prepare the low-level features as the complementary signals for the segmentor, since these structural landmarks cannot accurately determine the cardiac boundaries. Specifically, we apply the Canny [11] operator to both source and target images to acquire a wide range of edges in source and target images. Although Canny edges remove most of the domain specific visual information, we further add an discriminator D^c to encourage the encoder E^c to extract the domain invariant feature. The adversarial loss is defined as:

$$\begin{aligned} \mathcal{L}_{adv}^c = & \mathbb{E}_{X_s} \left[\log \left(D^c \left(E^c(\text{Canny}[x_s^1]) \right) \right) \right] \\ & + \mathbb{E}_{X_t} \left[\log \left(1 - D^c \left(E^c(\text{Canny}[x_t^1]) \right) \right) \right]. \quad (5) \end{aligned}$$

At last, the segmentor takes as input the high-level domain-invariant structural feature (represented by landmark-based Gaussian maps) and low-level features from Canny edges (obtained from both source and target domain images), to generate the final segmentation results. During the training phase, given that only the source domain is labeled, the network is optimized by minimizing the voxel-wise cross entropy loss \mathcal{L}_{seg}^s on the source domain data, while the forward segmentation results on target domain data is the *final prediction results* in testing phase.

C. Learning Process

The end-to-end framework is trained by minimizing three types of losses, including the reconstruction losses \mathcal{L}_{rec}^s (source domain) and \mathcal{L}_{rec}^t (target domain), the segmentation loss \mathcal{L}_{seg}^s (source domain), and the adversarial losses \mathcal{L}_{adv}^h and \mathcal{L}_{adv}^c . The overall objective function is as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{seg}^s + \mathcal{L}_{adv}^h + \mathcal{L}_{adv}^c + \lambda (\mathcal{L}_{rec}^s + \mathcal{L}_{rec}^t), \quad (6)$$

where λ is a balancing weight that is empirically set to 10.0 in all the experiments. As for the number of landmarks in the structure representation, we use $K = 32$ in the experiments and thoroughly validate it in the discussion section.

In the testing stage, instead of building data pairs, the testing image in the target domain directly goes through the 3D landmark detector, and the distilled landmark-based Gaussian maps are fed into the segmentor with corresponding Canny edges to generate the final segmentation results (the data flow is shown in Fig. 2 with green lines).

D. Network Configurations and Implementation Details

1) Network Backbone: The framework consists of 4 encoders $\{E^h, E^s, E^t, E^c\}$, 3 decoders $\{G^s, G^t, G\}$ and 2 discriminators $\{D^h, D^c\}$, which are all built on the 3D convolutional neural network. The detailed structure of the network

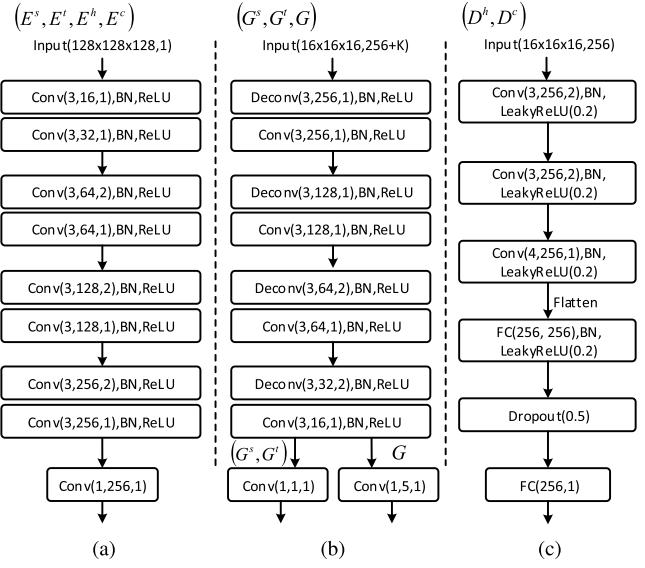


Fig. 3. The network details of the encoders, decoders and discriminators. “Input(a, c)” represents the size and channels of the input; “Conv/Deconv(k, n, s)” denotes the convolutional or deconvolutional layers with kernel size $k \times k \times k$, stride s and output channel n ; “FC($n1, n2$)” represents the fully connected layer with input channel $n1$ and output channel $n2$.

backbone is shown in Fig. 3. Note that the input of the decoders has the size of $(16 \times 16 \times 16)$ with $(256+K)$ channels, which is composed of 256 channels extracted by the encoders and K landmark-based Gaussian maps. Meanwhile, the output channel of the decoders $\{G^s, G^t\}$ is 1 for reconstruction, whereas G outputs 5 channels for cardiac segmentation, and each of them represents the class probability of being certain cardiac parts or the background.

2) Implementation Details: Our framework was implemented in Python with PyTorch platform. In training phase, four NVIDIA GTX 1080Ti GPUs are used to train the network in parallel. We employ Adam optimizer with a fixed learning rate of 1.0×10^{-3} . In total, we train the network for 40K iterations (about 16 hours) and the discriminators (D^h, D^c) are optimized every 10 iterations as is common in GAN network training.

IV. EXPERIMENTS

In this section, we evaluated our method on the public dataset, compared it against state-of-the-art methods and conducted comprehensive ablation studies to validate the effectiveness and accuracy. All experiments are performed on a machine with an Intel(R) Xeon(R) V4 1.9GHz CPU, 4 Nvidia 1080Ti GPUs, and 32GB RAM.

A. Dataset

The dataset used in this paper is the MICCAI 2017 Multi-Modality Whole Heart Segmentation (MM-WHS) [12] dataset, where 20 MRI and 20 CT scans with ground truth masks are provided for cardiac segmentation. Note that the images across modalities are unpaired, i.e., they are collected from different clinical sites and patients. In our experiments, four cardiac structures including the ascending aorta (AA), the left atrium blood cavity (LA-BC), the left

ventricle blood cavity (LV-BC), and the myocardium of the left ventricle (MYO) are adopted for the segmentation task.

As the CT and MR scans have different field of view, we manually crop the original scans to cover the heart region where we aim to segment, and resize all cropped scans with the size of $128 \times 128 \times 128$. Before inputting to the network, each sample image is normalized to have zero mean and unit variance in terms of the intensity value. We randomly split 20 target scans into two folds and perform the two-fold cross validation. In each fold, 20 scans of the source domain and 10 scans of the target domain are used to train the network, and the remaining 10 scans of the target domain are used for testing.

Remark: The data split scheme is commonly used by the state-of-the-art methods ([19], [42], [49]), we thus follow the same setup to do the fair evaluation and comparison. However, there are extra 40 non-annotated target domain data available in MM-WHS, which can be used for both training and testing in our pipeline. We further provide these results in the discussion section (Sec. V).

B. Experimental Settings and Evaluation Metrics

To conduct comprehensive analysis and comparisons, we design eight experimental settings, implement and test them on the MM-WHS dataset. For a fair comparison, in addition to the same datasets for training and testing, all the 3D version settings utilize the same backbone subnetworks, including the feature encoder, decoder and the segmentor.

- **W/o domain adaptation (WoDA):** applies the segmentor learned on the source domain to the target domain without domain adaptation, which achieves the lower bound performance. Note here, WoDA only consists of the basic encoder and decoder, without the landmark detection module and the adversarial training component.
- **Full supervision (FS):** trains our framework on both source and target domains with corresponding labels, which obtains the upper bound performance.
- **Feature adaptation (3D-ADDA):** extends ADDA [8] to the 3D version that aligns different modalities on feature space.
- **Image adaptation (3D-CycleGAN):** extends CycleGAN [5] to the 3D version that aligns different modalities on image space.
- **Feature and image adaptation (3D-CyCADA):** extends CyCADA [31] to the 3D version that aligns different modalities on both feature and image spaces.
- **Disentangled Representation (3D-DISE):** extends DISE [34] to the 3D version that distills the structure and texture features to adapt from virtual to real-world scenes, e.g. SYNTHIA [50].
- **SIFA [19]:** utilizes 2D neural network to conduct synergistic alignment on both image and feature spaces.
- **UESM [42]:** designs an uncertainty-aware domain adaptation network to boost the segmentation performance on highly uncertain regions.

To quantitatively evaluate the segmentation performance, we exploit two common metrics, the Dice similarity coefficient

(Dice) and the average symmetric surface distance (ASD), to measure the segmentation accuracy. Dice presents the voxel-wise segmentation accuracy, while ASD calculates the average surface distance between the predicted labels and its corresponding ground truth labels in 3D space. Note that a higher Dice value and a lower ASD value indicate higher quality in cardiac segmentation results. We present the metrics in the format of $mean \pm std$ to show the average performance as well as cross-subject variance.

C. Comparisons and Analysis

We evaluate our approach for bidirectional domain adaptation, i.e., from MRI to CT images and from CT to MRI images, respectively. To have a better understanding of the comparisons, we first establish the lower bound (WoDA) and upper bound (FS) network configurations in each direction. As shown in Table I, there are extreme performance degradations between the lower and upper bound neural models in both directions. Also pay attention to the extreme lower performance of WoDA from both directions that comes from the intrinsic domain gap between the two modalities. All these observations provide clear evidence of the necessity to propose unsupervised domain adaptation methods. Another interesting observation is that the lower bound of the adaptation direction from MRI to CT images has a significantly higher accuracy compared to the inverse direction in terms of the Dice score (64.9% vs. 38.1%). We hypothesize that it is due to the fact that MRI images have a relatively limited intensity contrast near the cardiac boundaries. The large intensity variation in CT images makes transferring the knowledge to MRI images more challenging, which is consistent to the performance of fully supervised learning (89.8% vs. 84.0%) and other domain adaptation methods ([5], [8], [19], [31], [31], [34], [42]) in both two adaptation directions.

To validate the effectiveness of our method, we compare with different state-of-the-art unsupervised domain adaptation approaches for natural images and the MM-WHS dataset. The quantitative comparisons are presented in Table I. Our method achieves the best performance in both directions by a large margin over other unsupervised methods, which shows the advantages of the anatomical structure representation in domain adaptation. In the meanwhile, 3D-ADDA and 3D-CycleGAN gain noticeable improvements compared to the model without domain adaptation (WoDA). Also, 3D-CycADA combining both feature- and image-level alignments further improves the segmentation performance over methods relying on either of them. Importantly, compared to 3D-DISE that distills structure and texture information in feature space, our method achieves favorable performance (8.2% and 9.6% improvements in terms of Dice score of two adaptation directions), demonstrating the effectiveness of our explicit structural landmarks to represent the human-body anatomical structure.

SIFA and UESM are two state-of-the-art unsupervised domain adaptation methods on the MM-WHS dataset. Note that, to have a fair comparison, we implement their methods, train and test them using our dataset split scheme, instead of their reported versions where 16 target data is used for

TABLE I
QUANTITATIVE RESULTS OF BIDIRECTIONAL DOMAIN ADAPTATION

| MRI to CT | | | | | | | | | | | |
|-----------------|-----|----------------------|----------------------|----------------------|----------------------|----------------------|---------------------|---------------------|---------------------|---------------------|---------------------|
| Methods | Dim | Dice [%] ↑ | | | | | ASD [Voxel] ↓ | | | | |
| | | AA | LA-BC | LV-BC | MYO | Mean | AA | LA-BC | LV-BC | MYO | Mean |
| WoDA FS | 3D | 63.6±4.2 90.3±1.4 | 61.8±4.3 90.1±1.5 | 70.9±3.4 91.1±1.5 | 63.3±4.7 86.7±1.7 | 64.9±4.2 89.8±1.5 | 16.3±4.0 2.4±1.1 | 32.2±6.7 2.9±1.3 | 17.0±3.8 2.4±0.3 | 19.1±4.0 2.1±0.7 | 21.2±4.9 2.5±0.8 |
| | 3D | 68.4±2.9 | 66.0±3.2 | 75.9±2.5 | 63.2±3.6 | 68.3±2.9 | 11.4±3.1 | 19.7±3.9 | 13.0±2.2 | 14.5±2.7 | 14.6±3.1 |
| 3D-ADDA [8] | 3D | 60.9±3.1 | 70.8±2.2 | 73.6±2.0 | 63.1±2.4 | 67.1±2.4 | 17.2±3.5 | 11.4±2.5 | 9.0±2.5 | 6.6±3.1 | 11.0±2.9 |
| 3D-CycleGAN [5] | 3D | 72.0±2.4 | 76.5±1.8 | 74.0±1.9 | 61.7±2.2 | 71.1±2.1 | 7.2±2.1 | 6.8±1.9 | 9.3±1.5 | 6.0±2.4 | 7.3±2.0 |
| 3D-CyCADA [32] | 3D | 81.2±1.9 | 80.1±1.5 | 78.5±1.8 | 70.6±2.1 | 77.6±1.8 | 7.0±0.7 | 6.5±0.5 | 7.5±2.6 | 6.5±1.8 | 6.9±1.7 |
| 3D-DISE [35] | 3D | 76.4±2.4 | 75.5±1.9 | 79.0±2.0 | 56.5±2.5 | 71.9±2.2 | 11.2±3.3 | 6.7±2.5 | 5.6±3.8 | 9.9±2.0 | 8.4±3.0 |
| SIFA [19] | 2D | 82.1±2.0 | 87.0±1.7 | 83.6±1.7 | 68.4±1.8 | 80.3±1.8 | 4.2±1.3 | 5.3±1.3 | 4.1±1.4 | 4.5±1.3 | 4.5±1.3 |
| UESM [43] | 2D | 87.9±1.7 | 88.1±1.5 | 88.4±1.7 | 78.7±2.0 | 85.8±1.6 | 3.8±0.6 | 3.3±1.2 | 3.1±1.2 | 3.4±1.2 | 3.4±1.0 |
| Ours | 3D | 87.9±1.7 | 88.1±1.5 | 88.4±1.7 | 78.7±2.0 | 85.8±1.6 | 3.8±0.6 | 3.3±1.2 | 3.1±1.2 | 3.4±1.2 | 3.4±1.0 |
| CT to MRI | | | | | | | | | | | |
| Methods | Dim | Dice [%] ↑ | | | | | ASD [Voxel] ↓ | | | | |
| | | AA | LA-BC | LV-BC | MYO | Mean | AA | LA-BC | LV-BC | MYO | Mean |
| WoDA FS | 3D | 22.5±5.6 77.7±2.0 | 39.5±4.6 82.6±1.6 | 55.2±2.5 93.6±1.5 | 35.3±5.1 82.1±1.8 | 38.1±4.5 84.0±1.7 | 25.1±4.8 1.7±1.1 | 17.9±3.9 2.2±1.2 | 15.9±5.0 2.0±0.5 | 10.6±4.2 1.6±0.2 | 17.4±4.4 1.9±0.8 |
| | 3D | 33.2±4.5 | 44.6±3.9 | 74.9±3.0 | 53.9±3.8 | 51.7±3.7 | 10.0±2.3 | 14.3±2.6 | 6.3±0.8 | 4.7±2.7 | 8.8±2.0 |
| 3D-ADDA [8] | 3D | 39.9±4.6 | 57.8±3.8 | 67.1±4.0 | 42.3±3.8 | 51.8±4.0 | 9.4±2.2 | 10.8±2.3 | 6.3±2.8 | 4.5±0.6 | 7.7±1.9 |
| 3D-CycleGAN [5] | 3D | 49.1±3.5 | 66.2±3.6 | 76.8±3.3 | 55.9±3.9 | 62.0±3.6 | 5.6±1.9 | 4.4±1.5 | 3.9±1.4 | 3.7±1.3 | 4.4±1.5 |
| 3D-CyCADA [32] | 3D | 66.2±2.6 | 63.0±3.0 | 73.1±2.8 | 58.6±3.1 | 65.2±2.8 | 5.4±1.7 | 5.0±1.3 | 4.1±0.5 | 4.7±2.2 | 4.8±1.3 |
| 3D-DISE [32] | 3D | 57.0±3.4 | 57.3±3.1 | 71.0±2.5 | 56.7±3.0 | 60.5±2.9 | 8.5±2.0 | 8.9±1.9 | 4.6±2.8 | 5.1±0.6 | 6.8±1.8 |
| SIFA [19] | 2D | 69.5±2.5 | 67.0±2.1 | 75.4±2.2 | 60.2±2.7 | 68.0±2.3 | 4.9±1.6 | 4.5±0.4 | 3.2±2.3 | 4.6±0.4 | 4.3±1.2 |
| UESM [43] | 2D | 72.8±1.7 | 79.3±1.5 | 82.3±1.8 | 64.7±1.9 | 74.8±1.7 | 2.2±0.5 | 2.8±1.4 | 2.8±1.2 | 2.4±0.5 | 2.6±0.9 |
| Ours | 3D | 72.8±1.7 | 79.3±1.5 | 82.3±1.8 | 64.7±1.9 | 74.8±1.7 | 2.2±0.5 | 2.8±1.4 | 2.8±1.2 | 2.4±0.5 | 2.6±0.9 |

training and only 4 data for testing. Compared to these two methods built on the 2D neural network, our framework based on the 3D neural network obtains better performance in both two adaptation directions. An interesting observation is that, UESM built on the 2D neural network, outperforms all the alternative methods with the 3D neural network, which is a bit counterintuitive and can be explained from two aspects. First, other than the 3D lower bound network WoDA, we build the 2D lower bound network WoDA-2D, which achieves 46.8% and 28.7% mean Dice accuracy for MRI to CT and CT to MRI directions, respectively. Compared to WoDA (64.9% and 38.1%), the performance is much lower and consistent with the consensus that 3D networks outperforms the 2D counterpart. Second, the favorable result of UESM mainly comes from the special designed uncertainty-aware self-training strategy and feature recalibration module that boost the performance and play the key roles.

The representative visual segmentation examples are presented in Figs. 4 and 5, where our approach can accurately segment the four cardiac substructures compared to the ground truth (the second column), while other methods either fail to produce the segmentation or generate inaccurate boundaries. Meanwhile, the reconstructed high-quality 3D model of four cardiac structures (Fig. 5) shows the consistent segmentation accuracy as reflected in the statistics. More importantly, as shown in Fig. 5, the learned 3D landmarks lying in the cardiac area (overlaid with the 3D model) extract consistent local features within each adaptation direction, demonstrating the network has learned semantically meaningful keypoints.

1) Topological Consistency of Landmarks: The visual results in Fig. 5 reflect the topological consistency of the landmarks to some extent, we further quantify it using the normalized distances between points as the metric. Specifically, we have

10 cardiac testing data with ground truth segmentation masks and we first align them together. Suppose we take the 1-st sample as a reference, and apply affine and deformable transformations to any of the remaining 9 samples to align them to the reference. The transformations are calculated from the ground truth masks and then applied to the corresponding landmarks. After the alignment, we calculate the average normalised distance between points as the metric. Since we have 10 samples, each one will serve as the reference one time, and we average the 10 values as the final topological consistency metric. In our test dataset, the topological consistency distance is 4.6 voxel-size (about 2.5mm), which is much smaller compared to the whole cardiac region size (about 200mm).

2) Comparison With Traditional Method: [51] proposes the state-of-the-art traditional method for whole heart segmentation using non-linear registration and non-local fusion. However, this method is designed for single-modality adaptation. To achieve the fair comparison, we re-implement it to perform multi-modality adaptation in the following.

As shown in Fig. 6, given an MRI image (Fig. 6(a)), we first applied the affine and deformable transformations to align it to the target CT image (Fig. 6(b)), so that we get a deformed MRI image (Fig. 6(c)). Then, the same transformations are applied to the ground truth segmentation mask of the MRI image to derive the resulting segmentation mask of the target CT image (Fig. 6(d)). As marked by the red arrow, we can see clearly that the resulting segmentation does not respect the shape boundary very well, while our method achieves the accurate result (Fig. 6(e)). We further tested the segmentation accuracy of the MRI to CT adaptation results of [51], and it achieved 67.07% mean Dice accuracy, while our mean Dice accuracy is 85.8% that outperforms [51] by a large margin. Notably,

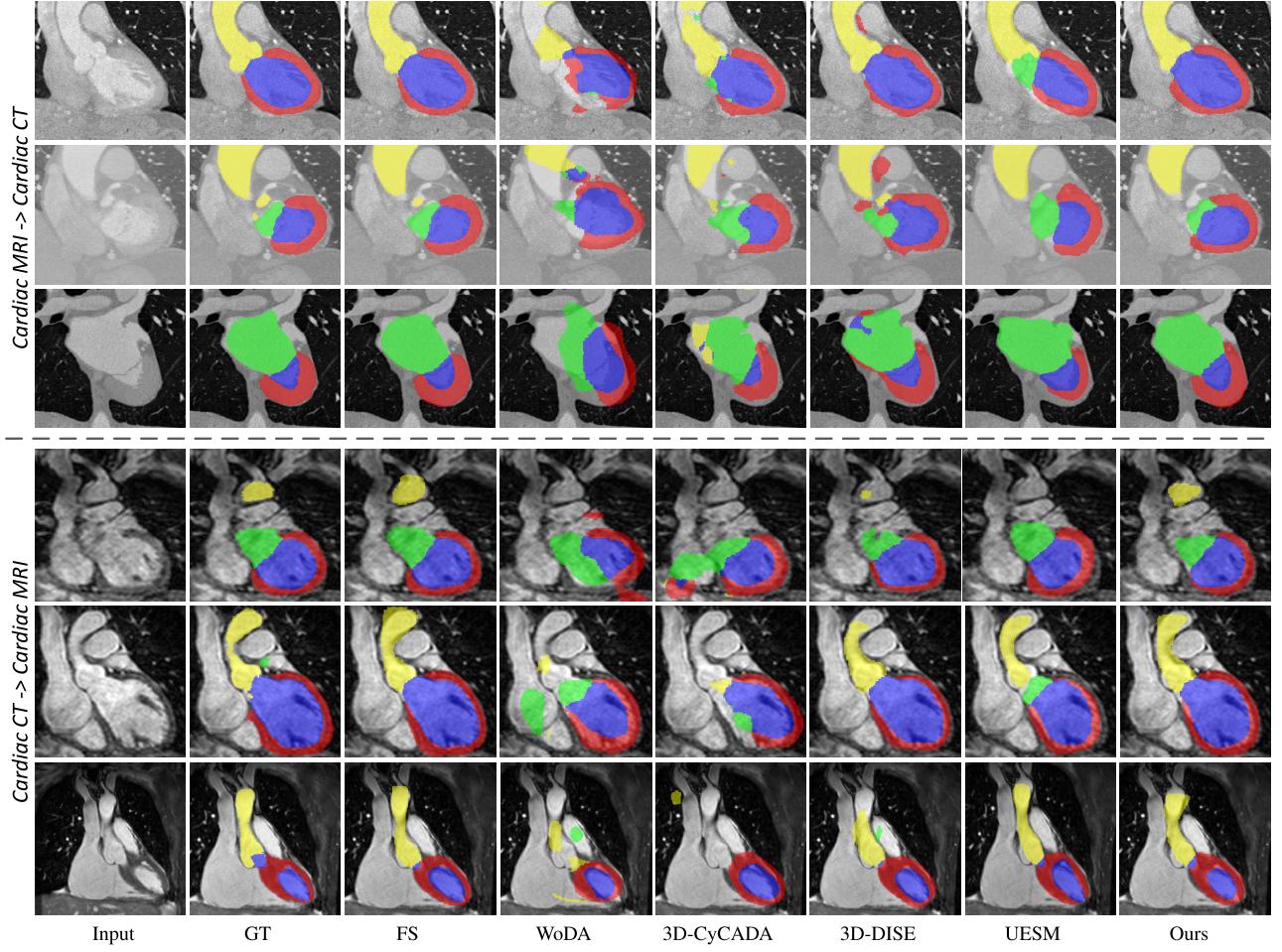


Fig. 4. Qualitative results of different methods for unsupervised MRI to CT domain adaptation (top three rows) and CT to MRI domain adaptation (bottom three rows). Typical examples are showed row-by-row. The yellow, green, red and blue colors represent the cardiac structures AA, LA-BC, LV-BC and MYO, respectively.

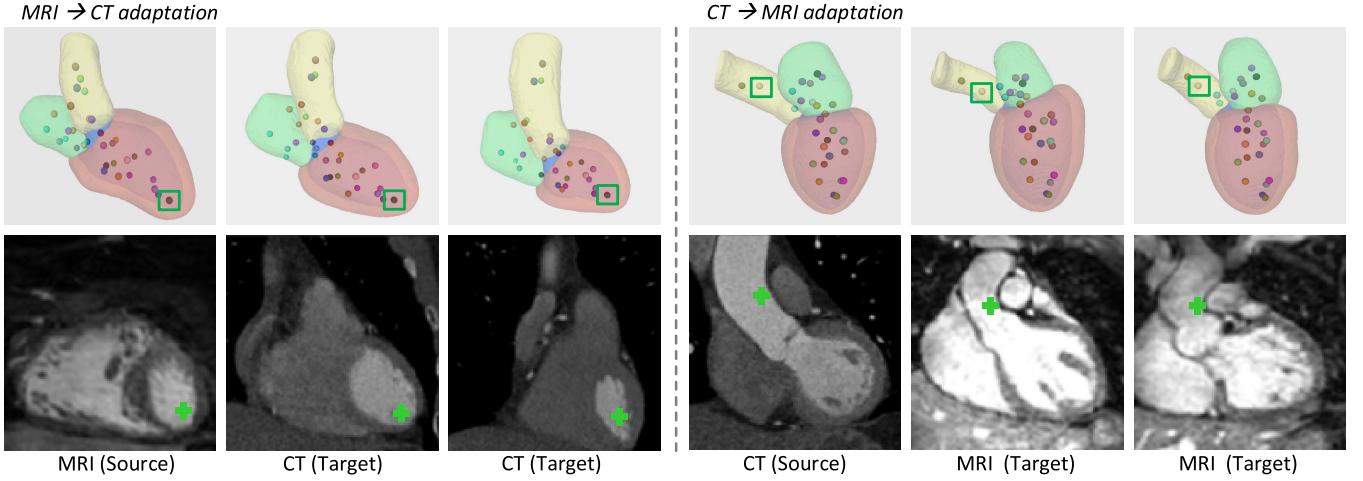


Fig. 5. 3D segmentation results and corresponding extracted landmarks, including MRI to CT adaptation (left) and CT to MRI adaptation (right). The first row overlays the segmented cardiac sub-structures with spatial landmarks, where the color-coding of landmarks expresses the consistency across modalities. The second row visualizes the 2D coronal view image slice on the location of the selected landmark point (green box). Each column corresponds to one example.

all learning-based methods reported in our paper obtain better performance than [51].

3) Paired t-test With Other Methods: we have computed the p-value using paired t-tests to illustrate the significance of

our performance improvements over other methods. We utilized Dice and ASD as the evaluation metrics and set the significance level as 0.05. As shown in Table II, all paired t-tests in both adaptation directions present p-value smaller

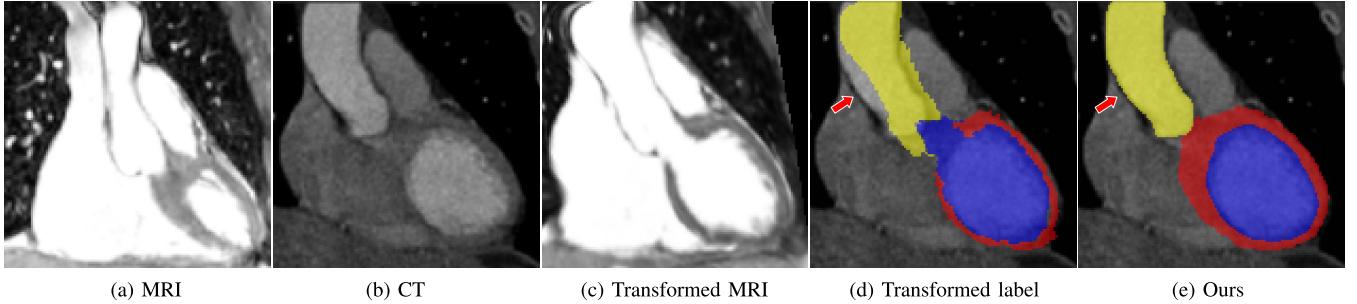


Fig. 6. The domain adaptation results from [4]. (a) The source MRI image, (b) the target CT image, (c) the deformed MRI image after registration to (b), (d) the resulting segmentation result of (b) by adapting the ground truth segmentation mask from (a), (e) the segmentation result of our method.

TABLE II
PAIRED T-TEST FOR OUR METHOD WITH OTHER METHODS

| | Metric | 3D-ADDA | 3D-CycleGAN | 3D-CyCADA | 3D-DISE | SIFA | UESM |
|-----------|--------|---------|-------------|-----------|---------|------|------|
| MRI to CT | Dice | 2e-10 | 2e-8 | 7e-10 | 3e-6 | 9e-9 | 1e-7 |
| | ASD | 8e-10 | 5e-6 | 6e-6 | 7e-6 | 5e-7 | 4e-6 |
| CT to MRI | Dice | 3e-10 | 4e-10 | 4e-8 | 8e-6 | 3e-8 | 7e-7 |
| | ASD | 4e-9 | 7e-8 | 5e-5 | 2e-4 | 2e-6 | 4e-6 |

than 0.05, indicating that our performance improvements over other methods are statistically significant.

D. Ablation Study of Key Components

Technically, we have three key components used in our framework, including the unsupervised landmark detection, Canny edge feature extraction, and discriminators for adversarial training. We have designed a full ablation study with all the combinations of the ingredients. Since the adversarial training is coupled with Canny edge or landmarks, we thus have the following configurations:

- Canny: it only has a segmentor trained on the Canny edges extracted from source domain data and directly forwards the Canny edges from the target domain data to get the resulting segmentation.
- Landmark: it has the unsupervised landmark detection module, and in the segmentation module, the segmentor only accepts the Gaussian map from landmarks as input to derive the resulting segmentation.
- Canny-Adv: augment Canny with the discriminator D^c , so that the segmentor takes as input the Canny edges extracted from both the source and target domains.
- Landmark-Adv: augment Landmark with the discriminator D^h to further align the features.
- Landmark-Canny: augment Landmark with the Canny edge input, so that the segmentor takes both the Gaussian map from the landmarks and the Canny edge features.
- FullNet (Landmark-Canny-Adv): combine all the three ingredients together.

We also include the lower bound WoDA (without any of the three ingredients) to better measure the improvements of different ingredients. The statistics for the bidirectional adaptations are presented in Table III, where the three components work together (FullNet) to achieve the best performance, except that the ASD error of the AA structure in the MRI to CT direction outperforms a little bit due to the variability and the inherent stochasticity of network training. We discuss each component in the following:

1) Effectiveness of Landmarks: The detected landmark that encodes the underlying domain-invariant anatomical structure is the key factor of our algorithm, and is also the reason why our method can outperform the state-of-the-art approaches. This can be validated from two aspects. Firstly, all configurations with landmarks, e.g., Landmark, Landmark-Adv, achieve higher segmentation results and outperform the state-of-the-art methods in both adaptation directions. Secondly, the performance improvement also reveals the conclusion. Take the MRI to CT adaptation as an example, compared to WoDA, Landmark and Canny boost the Dice accuracy by 16.4% and 9.5%, respectively. And adding Canny to Landmark only gets a 1.6% Dice increase, while adding Landmark to Canny obtains a remarkably 8.5% increase. Similarly, by augmenting Canny to Landmark-Adv, the Dice accuracy grows only 2.3%, while by augmenting Landmark to Canny-Adv, the Dice accuracy rises by a considerable 6.2%.

The similar statistical evidence can be derived from the CT to MRI adaptation direction, and all statistical analyses explain the conclusion.

2) Effectiveness of Canny Edges: Canny edge map is effective and helps significantly, especially for the MRI to CT adaptation. For example, Canny boosts the performance of WoDA by about 9.5%, and outperforms most of the state-of-the-art methods. However, it cannot lead to the favorable performance for the CT to MRI adaptation, where Canny only gains 2.6% improvement over WoDA and it is lower than all the state-of-the-art methods. The reason is that CT images display more clear cardiac structure than MRI images, thus the Canny edge map extracted from CT images is less noisy and more helpful. A well-trained segmentor on the clear Canny edges performs worse when facing noisy input (CT to MRI), while it obtains good results in the inverse direction since CT Canny edges have inherent clear structure (MRI to CT).

3) Effectiveness of Adversarial Training: The adversarial training serves to further align the source and target domain data in Canny edge feature extraction and landmark detection. As listed in Table III, by adding adversarial training, Canny-Adv, Landmark-Adv, FullNet improve the mean Dice accuracy by 5.2%, 2.2%, and 2.7% for the MRI to CT adaptation, while they gains 2.6%, 2.8%, and 1.7% improvement for the CT to MRI adaptation. In addition, a notable and interesting observation is that without D^h and D^c , all the configurations still achieve promising segmentation results, e.g., for the MRI to CT adaptation, Landmark-Canny achieves 82.9% on

TABLE III
ABLATION STUDIES OF THE NETWORK KEY COMPONENTS FOR BOTH ADAPTATION DIRECTIONS

| Methods | MRI to CT | | | | | ASD [Voxel] ↓ | | | | |
|----------------|-----------------|-----------------|-----------------|-----------------|-----------------|---------------|----------------|----------------|----------------|----------------|
| | AA | LA-BC | LV-BC | MYO | Mean | AA | LA-BC | LV-BC | MYO | Mean |
| WoDA | 63.6±4.2 | 61.8±4.3 | 70.9±3.4 | 63.3±4.7 | 64.9±4.2 | 16.3±4.0 | 32.2±6.7 | 17.0±3.8 | 19.1±4.0 | 21.2±4.9 |
| Canny | 80.6±3.5 | 78.2±3.8 | 73.3±2.5 | 65.7±4.9 | 74.4±3.7 | 8.5±2.8 | 7.7±2.5 | 5.1±2.2 | 6.7±2.4 | 7.0±2.5 |
| Landmark | 84.4±1.9 | 82.0±1.2 | 83.6±2.0 | 75.1±3.3 | 81.3±2.0 | 6.8±2.3 | 5.2±1.8 | 4.8±1.6 | 4.7±0.8 | 5.4±1.6 |
| Canny-Adv | 85.1±2.1 | 81.2±1.8 | 78.1±3.4 | 75.2±2.0 | 79.6±2.3 | 4.9±1.1 | 5.1±1.6 | 7.1±2.0 | 5.5±0.7 | 5.6±1.4 |
| Landmark-Adv | 86.4±2.0 | 83.6±2.1 | 87.5±1.6 | 76.8±1.8 | 83.5±1.8 | 7.8±1.9 | 4.8±1.1 | 3.1±0.7 | 3.5±0.8 | 4.8±1.1 |
| Landmark-Canny | 86.4±1.5 | 82.5±1.4 | 85.2±1.2 | 77.2±2.5 | 82.9±1.7 | 3.5±0.8 | 5.2±1.9 | 4.4±1.3 | 3.9±1.4 | 4.2±1.3 |
| FullNet | 87.9±1.7 | 88.1±1.5 | 88.4±1.7 | 78.7±2.0 | 85.8±1.6 | 3.8±0.6 | 3.3±1.2 | 3.1±1.2 | 3.4±1.2 | 3.4±1.0 |

| Methods | CT to MRI | | | | | ASD [Voxel] ↓ | | | | |
|----------------|-----------------|-----------------|-----------------|-----------------|-----------------|----------------|----------------|----------------|----------------|----------------|
| | AA | LA-BC | LV-BC | MYO | Mean | AA | LA-BC | LV-BC | MYO | Mean |
| WoDA | 22.5±5.6 | 39.5±4.6 | 55.2±2.5 | 35.3±5.1 | 38.1±4.5 | 25.1±4.8 | 17.9±3.9 | 15.9±5.0 | 10.6±4.2 | 17.4±4.4 |
| Canny | 30.8±7.1 | 42.7±5.2 | 52.4±4.4 | 36.9±4.4 | 40.7±5.0 | 16.9±4.0 | 12.8±3.2 | 17.0±2.2 | 10.3±3.9 | 14.3±3.3 |
| Landmark | 66.9±3.0 | 72.8±2.8 | 79.0±2.5 | 59.7±2.6 | 69.6±2.7 | 5.9±1.7 | 6.8±1.7 | 6.1±2.5 | 5.3±0.8 | 6.0±1.6 |
| Canny-Adv | 33.8±6.3 | 45.5±4.8 | 56.2±5.1 | 37.5±4.5 | 43.3±4.9 | 12.4±4.2 | 9.7±2.9 | 14.3±2.9 | 8.4±1.7 | 11.2±2.9 |
| Landmark-Adv | 70.1±2.2 | 76.9±1.8 | 80.8±1.7 | 61.6±2.7 | 72.4±2.1 | 4.5±1.5 | 5.3±1.4 | 5.1±1.8 | 4.7±0.8 | 4.9±1.3 |
| Landmark-Canny | 70.2±1.6 | 78.9±1.5 | 80.4±1.7 | 62.9±2.1 | 73.1±1.6 | 3.1±1.4 | 3.6±1.3 | 2.8±0.7 | 3.2±1.7 | 3.2±1.3 |
| FullNet | 72.8±1.7 | 79.3±1.5 | 82.3±1.8 | 64.7±1.9 | 74.8±1.7 | 2.2±0.5 | 2.8±1.4 | 2.8±1.2 | 2.4±0.5 | 2.6±0.9 |

average DSC score and 4.2 on the average ASD error, which is much different with the severe failure without adversarial training observed in previous works [7], [19]. This also verifies our strategy design that domain-invariant structural landmarks as well as Canny edges play important roles in unsupervised domain adaptation.

4) Paired t-test With Different Ablation Configurations: Similarly, we have conducted paired t-tests for our FullNet with other ablation configurations. It can be observed from Table IV, all the p-values are smaller than 0.05, except the mean ASD of ‘Landmark-Canny’ from the CT to MRI adaptation direction (i.e., 0.058). This is because by exploiting landmarks and Canny edges, the network version ‘Landmark-Canny’ has already achieved good performance, thus by further adding adversarial training (i.e., FullNet) would gain a little improvement, which is not that significant. This also reveals that landmarks and Canny edges play more important roles in our pipeline.

E. Sensitivity of the Canny Edge Noise

Generally, Canny edges contain a certain level of noises and imperfections (e.g., open-corners or missed junctions) by applying different parameter threshold values (i.e., the deviation of the Gaussian filter in the Canny operator). However, our method is robust to the noises and imperfections, because the low-level appearance information encoded in the edge map, instead of the exact corners and junctions, is more practical in our case, since we find the segmentation boundary by regression instead of sparse line selection. To validate the sensitivity of our method to noises or imperfections, we have conducted a ablation study by varying the deviation values.

Qualitatively, as shown in Fig. 7, both Canny edges with increasing noises and imperfections lead to similar accurate segmentation results compared to the ground truth. And quantitatively (Table V), all the configurations achieved comparable segmentation quality in terms of the mean Dice accuracy

TABLE IV
PAIRED T-TEST FOR OUR FULLNET WITH OTHER
ABLATION CONFIGURATIONS

| | Metric | WoDA | Canny | Landmark | Canny-Adv | Landmark- -Adv | Landmark- -Canny |
|-----------|--------|-------|-------|----------|-----------|-------------------|---------------------|
| MRI to CT | Dice | 2e-10 | 1e-8 | 5e-6 | 3e-7 | 6e-5 | 3e-5 |
| | ASD | 6e-9 | 6e-8 | 4e-6 | 8e-7 | 3e-4 | 5e-4 |
| CT to MRI | Dice | 1e-11 | 4e-11 | 3e-6 | 1e-10 | 3e-5 | 5e-3 |
| | ASD | 5e-10 | 8e-10 | 7e-5 | 8e-9 | 6e-3 | 0.058 |

TABLE V
STATISTIC RESULTS OF THE SEGMENTATION QUALITY USING
DIFFERENCE CANNY EDGES

| Dice [%] | Deviation | | | | |
|-----------|-----------|----------|----------|----------|----------|
| | 1 | 2 | 3 | 4 | 5 |
| MRI to CT | 85.0±1.6 | 85.6±1.8 | 85.8±1.6 | 85.8±1.6 | 85.7±1.9 |
| CT to MRI | 73.1±2.6 | 74.7±1.4 | 74.8±1.7 | 74.4±2.7 | 74.6±0.9 |

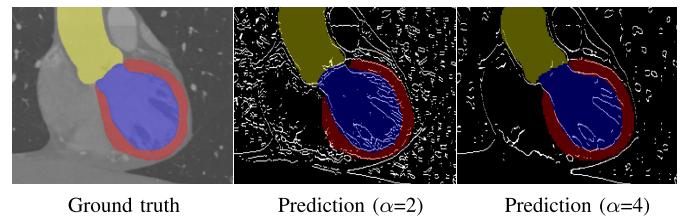


Fig. 7. Segmentation results by inputting different edge maps derived with different deviations of the gaussian filter.

in both adaptation directions. Both the visual and statistical results verify that our method is robust to Canny edge noises to generate high-quality segmentation results. As for the time efficiency, it takes about 2.1s to calculate a 3D Canny edge map, while the overall running time for a data sample in the testing stage is about 3.2s. Although we cannot achieve a real-time feedback, it is relatively fast enough.

F. Sensitivity of the Thin-Plate-Spline Deformation

We use the 3D thin-plate-spline (TPS) geometric deformation in our method to build the image pairs. To validate

TABLE VI

SEGMENTATION PERFORMANCE USING DIFFERENCE PARAMETERS OF TPS DEFORMATION IN THE IMAGE PAIR BUILDING STAGE

| R_s | 0.3 | 0.5 | 0.7 | 0.9 |
|-----------|----------|----------|----------|----------|
| Dice [%] | 84.9±1.8 | 85.8±2.0 | 85.8±1.6 | 82.4±1.6 |
| Nb_{pt} | 1x1x1 | 2x2x2 | 3x3x3 | 4x4x4 |
| Dice [%] | 82.8±2.1 | 84.9±1.5 | 85.8±1.6 | 85.6±1.7 |

TABLE VII

SEGMENTATION PERFORMANCE (MRI TO CT) OF OUR METHOD WITH DIFFERENT NUMBERS OF LANDMARKS

| # Landmarks | Dice [%] | | | | |
|-------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| | AA | LA-BC | LV-BC | MYO | Mean |
| 16 | 86.3±2.5 | 84.2±2.0 | 82.4±2.0 | 72.5±1.2 | 81.4±1.7 |
| 32 | 87.9±1.7 | 88.1±1.5 | 88.4±1.7 | 78.7±2.0 | 85.8±1.6 |
| 48 | 85.8±1.6 | 88.9±1.4 | 85.0±2.5 | 78.2±1.1 | 84.5±1.6 |
| 64 | 88.0±2.6 | 88.5±2.7 | 85.5±2.2 | 77.7±2.6 | 85.0±2.1 |

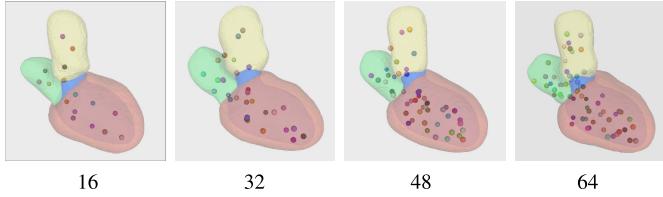


Fig. 8. Segmentation performance (MRI to CT) of our method with different numbers of landmarks.

the sensitivity to the deformation field, we do a further ablation experiment by changing two core parameters of TPS deformation: the deformation span range (R_s), and the number of points in the control grid (Nb_{pt}). Note that when we change one parameter, we use a default value for the other, i.e., 0.7 for the deformation span range, and $3 \times 3 \times 3 = 27$ for the number of points in the control grid, and report the mean Dice accuracy of the segmentation in Table VI.

As observed, except for two extreme cases (i.e., $R_s = 0.9$ and $Nb_{pt} = 1 \times 1 \times 1$), all other configurations achieve comparable high-quality segmentation results, which indicates that our method is insensitive to the deformation fields within the normal parameter range.

G. Sensitivity to the Number of 3D Landmarks

The 3D landmarks are learned unsupervisedly without manual annotation. To validate the performance of our method with different numbers of landmarks, we follow the same two-fold cross validation training strategies with 16, 32, 48 and 64 landmarks respectively. As shown in Table VII, the segmentation results using 16 landmarks are relatively lower compared to others, which indicates that it is insufficient to represent the cardiac structure properly with only 16 landmarks. Meanwhile, the comparable results achieved with 32, 48, and 64 landmarks reveal that our method is less sensitive to the number of landmarks within a certain range. In addition, we also provide the 3D segmentation results and corresponding landmarks in Fig. 8. It can be seen that, although there are different numbers of landmarks, they tend to represent the cardiac structure with a similar spatial distribution.

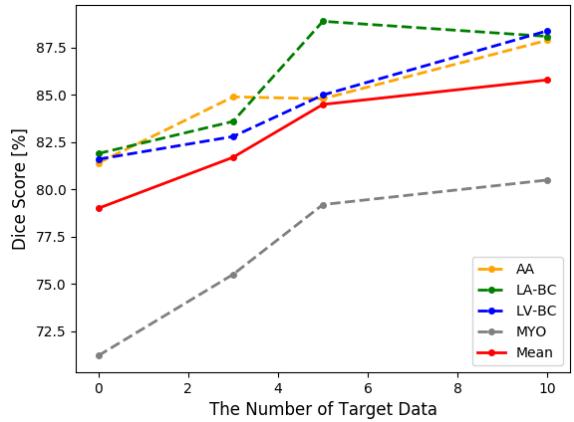


Fig. 9. Segmentation performance (MRI to CT) of our method with different numbers of target domain training scans.

H. Sensitivity to Target Domain Training Size

Unsupervised domain adaptation with limited data is also an important direction in medical image analysis [52]. To study the data efficiency of our method, we follow the same experiment settings except the number of target domain training scans. Perhaps surprisingly, as shown in Fig. 9, even without target domain training data, our method does not suffer from huge performance degradation and still outperforms the previous methods relying on image- and feature-level alignments. This is benefited from the fact that the learned cardiac structure and Canny edges are less effected by the domain-specific information. Meanwhile, more target domain scans seen by our framework will consistently improve the segmentation performance.

V. DISCUSSION

In contrast to most existing methods [7], [19] that focus on image- and feature-level alignments, to address the unsupervised domain adaptation problem, we explore a novel direction that utilizes 3D landmarks to represent the cardiac structure and combine it with low-level Canny edges to provide an accurate cardiac segmentation. To fully understand the mechanism of our method, we present the following discussions from two aspects, the advantages of explicit structure learning over implicit feature adaptation in segmentation tasks and exploiting the extra target domain data.

1) *Feature Adaptation vs. Anatomic Structure Detection:* We have claimed in the introduction that styles or latent features of images may not guarantee good domain adaptation results, which is more clear after results. This conclusion mainly comes from the comparison, the detailed ablation study and the discussion sections, where we clearly find that: 1) the state-of-the-art methods relying on style or latent feature adaptation achieve much lower segmentation quality; 2) our favorable segmentation quality is primarily due to the use of the automatically detected domain-invariant anatomical structure.

Thus, we conclude that the implicit common features existing in the image style or deep features across images of different modalities are less effective than the intrinsic common anatomical structure, since medical images of

TABLE VIII

THE STATISTIC OF THE ADDITIONAL EXPERIMENTS USING THE EXTRA 40 NON-ANNOTATED TARGET DOMAIN DATA

| | Dice [%] | | | | |
|-----------|----------|----------|----------|----------|----------|
| | AA | LA-BC | LV-BC | MYO | Mean |
| FullNet | 87.9±1.7 | 88.1±1.5 | 88.4±1.7 | 78.7±2.0 | 85.8±1.6 |
| FullNet-E | 87.4±2.5 | 87.6±2.3 | 88.0±2.1 | 78.8±2.6 | 85.5±2.3 |
| FullNet-T | 88.5±0.8 | 88.1±1.0 | 88.6±1.2 | 79.0±1.0 | 86.1±0.9 |

different modalities are captured to reveal the same anatomical structures. Thus explicitly constraining the neural models with respect to such common anatomical structures across images of different modalities benefit a lot of the domain adaptation.

2) Exploit Extra Target Domain Data in MM-WHS: there are extra 40 non-annotated target domain data available in MM-WHS, we have further explored to use these data in our algorithm for training and testing.

Training: since there is no requirement of ground truth labels from the target domain data in both landmark detection and segmentation modules, we thus trained our FullNet from scratch using 20 source domain data, 10 target domain data (the common splitting scheme), as well as the new 40 target domain data. The new network version is denoted as FullNet-T, and we reported the statistics on the same testing dataset (10 target domain data) in Table VIII. As can be seen, there is only a slight performance improvement (0.3%) with more target domain data. Combining this observation with the result in Fig. 9, it suggests that 10 target domain data is sufficient to robustly learn the domain-invariant landmarks, which plays the key role in the segmentation task.

Testing: we directly tested our trained FullNet on the 40 target domain data (the new testing version is denoted as FullNet-E). To be more clear, the 40 target domain data is actually labeled, but the labels are inaccessible and we only can run the special designed script to get the evaluation metrics. As can be seen from Table VIII, the mean Dice accuracy on the new testing dataset is 85.5%, which is comparable with FullNet (85.8%, tested on the 10 target domain dataset), indicating that our reported results are not cherry-picked and our method is robust for the domain adaptation task.

3) Limitation and Future Work: Although our approach has achieved good performance in bidirectional cardiac images (CT and MRI) adaptation, the limitation of our method still exists. Specifically, we assume both source and target data have relative high resolutions to capture the 3D anatomical structure. However, in clinical applications, the slice thickness of CT and MRI images may differ greatly, which brings new challenges to extract the meaningful and consistent 3D structural landmarks. Therefore, in the future, we would explore efficient approaches to combine 2D and 3D anatomical information for unsupervised domain adaptation. Meanwhile, since the anatomical structure is learned unsupervised without manual annotation, another interesting and appealing direction is to apply our unsupervised domain adaptation module in semi-supervised tasks so that the learned structural landmarks embedded in both labeled and unlabeled data would boost the performance in medical image processing tasks.

VI. CONCLUSION

We propose a novel structure-driven domain adaptation approach for unsupervised cross-modality cardiac segmentation. Our framework explicitly extracts the domain-invariant anatomical structure represented by 3D landmarks and combine it with the edge information to guide the accurate cardiac segmentation. We have evaluated our algorithm both qualitatively and quantitatively, and compared it against state-of-the-art methods, where our approach produces superior results and outperforms others by a significant margin. In addition, our proposed method is a general strategy that could be extended to other unsupervised domain adaptation tasks in medical image analysis.

REFERENCES

- [1] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.
- [2] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, “3D U-Net: Learning dense volumetric segmentation from sparse annotation,” in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2016, pp. 424–432.
- [3] Q. Dou *et al.*, “3D deeply supervised network for automated segmentation of volumetric medical images,” *Med. Image Anal.*, vol. 41, pp. 40–54, Oct. 2017.
- [4] I. Goodfellow *et al.*, “Generative adversarial nets,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.
- [5] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2223–2232.
- [6] Y. Huo *et al.*, “SynSeg-Net: Synthetic segmentation without target modality ground truth,” *IEEE Trans. Med. Imag.*, vol. 38, no. 4, pp. 1016–1025, Apr. 2019.
- [7] Q. Dou *et al.*, “PnP-AdaNet: Plug-and-play adversarial domain adaptation network at unpaired cross-modality cardiac segmentation,” *IEEE Access*, vol. 7, pp. 99065–99076, 2019.
- [8] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, “Adversarial discriminative domain adaptation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2962–2971.
- [9] Y. Zhang, Y. Guo, Y. Jin, Y. Luo, Z. He, and H. Lee, “Unsupervised discovery of object landmarks as structural representations,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2694–2703.
- [10] T. Jakab, A. Gupta, H. Bilen, and A. Vedaldi, “Unsupervised learning of object landmarks through conditional image generation,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 4016–4027.
- [11] J. Cannby, “A computational approach to edge detection,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-8, no. 6, pp. 679–698, Nov. 1986.
- [12] X. Zhuang and J. Shen, “Multi-scale patch and multi-modality atlases for whole heart segmentation of MRI,” *Med. Image Anal.*, vol. 31, pp. 77–87, Jul. 2016.
- [13] L. Wang, H.-M. Lai, G. J. Barker, D. H. Miller, and P. S. Tofts, “Correction for variations in MRI scanner sensitivity in brain studies with histogram matching,” *Magn. Reson. Med.*, vol. 39, no. 2, pp. 322–327, Feb. 1998.
- [14] L. G. Nyúl and J. K. Udupa, “On standardizing the MR image intensity scale,” *Magn. Reson. Med.*, vol. 42, no. 6, pp. 1072–1081, Dec. 1999.
- [15] T. Heimann, P. Mountney, M. John, and R. Ionasec, “Learning without labeling: Domain adaptation for ultrasound transducer localization,” in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Berlin, Germany: Springer, 2013, pp. 49–56.
- [16] R. Bermúdez-Chacón, C. Becker, M. Salzmann, and P. Fua, “Scalable unsupervised domain adaptation for electron microscopy,” in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2016, pp. 326–334.
- [17] A. van Opbroek, M. A. Ikram, M. W. Vernooij, and M. de Bruijne, “Transfer learning improves supervised image segmentation across imaging protocols,” *IEEE Trans. Med. Imag.*, vol. 34, no. 5, pp. 1018–1030, May 2015.

- [18] M. Ghafoorian *et al.*, “Transfer learning for domain adaptation in MRI: Application in brain lesion segmentation,” in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2017, pp. 516–524.
- [19] C. Chen, Q. Dou, H. Chen, J. Qin, and P. A. Heng, “Unsupervised bidirectional cross-modality adaptation via deeply synergistic image and feature alignment for medical image segmentation,” *IEEE Trans. Med. Imag.*, vol. 39, no. 7, pp. 2494–2505, Jul. 2020.
- [20] M. A. Degel, N. Navab, and S. Albarqouni, “Domain and geometry agnostic CNNs for left atrium segmentation in 3D ultrasound,” in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2018, pp. 630–637.
- [21] J. Ren, I. Hacihaliloglu, E. A. Singer, D. J. Foran, and X. Qi, “Adversarial domain adaptation for classification of prostate histopathology whole-slide images,” in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2018, pp. 201–209.
- [22] L. Zhang, M. Pereañez, S. K. Piechnik, S. Neubauer, S. E. Petersen, and F. A. Frangi, “Multi-input and dataset-invariant adversarial learning (MDAL) for left and right-ventricular coverage estimation in cardiac MRI,” in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2018, pp. 481–489.
- [23] S. Wang, L. Yu, X. Yang, C.-W. Fu, and P.-A. Heng, “Patch-based output space adversarial learning for joint optic disc and cup segmentation,” *IEEE Trans. Med. Imag.*, vol. 38, no. 11, pp. 2485–2495, Nov. 2019.
- [24] T. Joyce, A. Chartsias, and S. A. Tsaftaris, “Deep multi-class segmentation without ground-truth labels,” in *Proc. Int. Conf. Med. Imag. With Deep Learn.*, 2018, pp. 1–9.
- [25] N. Dong, M. Kampffmeyer, X. Liang, Z. Wang, W. Dai, and E. Xing, “Unsupervised domain adaptation for automatic estimation of cardio-thoracic ratio,” in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2018, pp. 544–552.
- [26] Y. Zhang, S. Miao, T. Mansi, and R. Liao, “Task driven generative modeling for unsupervised domain adaptation: Application to X-ray image segmentation,” in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2018, pp. 599–607.
- [27] C. Chen, Q. Dou, H. Chen, and P.-A. Heng, “Semantic-aware generative adversarial nets for unsupervised domain adaptation in chest X-ray segmentation,” in *Proc. Int. Workshop Mach. Learn. Med. Imag.*, 2018, pp. 143–151.
- [28] J. Jiang *et al.*, “Tumor-aware, adversarial domain adaptation from CT to MRI for lung cancer segmentation,” in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2018, pp. 777–785.
- [29] Z. Zhang, L. Yang, and Y. Zheng, “Translating and segmenting multimodal medical volumes with cycle- and shape-consistency generative adversarial network,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9242–9251.
- [30] C. Chen, Q. Dou, H. Chen, J. Qin, and P.-A. Heng, “Synergistic image and feature adaptation: Towards cross-modality domain adaptation for medical image segmentation,” in *Proc. 33rd AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 865–872.
- [31] J. Hoffman *et al.*, “Cycada: Cycle-consistent adversarial domain adaptation,” in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1994–2003.
- [32] Y. Zhang, Z. Qiu, T. Yao, D. Liu, and T. Mei, “Fully convolutional adaptation networks for semantic segmentation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6810–6818.
- [33] K. Kamnitsas *et al.*, “Unsupervised domain adaptation in brain lesion segmentation with adversarial networks,” in *Proc. Int. Conf. Inf. Process. Med. Imag.* Cham, Switzerland: Springer, 2017, pp. 597–609.
- [34] W.-L. Chang, H.-P. Wang, W.-H. Peng, and W.-C. Chiu, “All about structure: Adapting structural information across domains for boosting semantic segmentation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1900–1909.
- [35] L. Zhang, G.-J. Qi, L. Wang, and J. Luo, “AET vs. AED: Unsupervised representation learning by auto-encoding transformations rather than data,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2547–2555.
- [36] X. Wang, D. Kihara, J. Luo, and G.-J. Qi, “EnAET: A self-trained framework for semi-supervised and supervised learning with ensemble transformations,” 2019, *arXiv:1911.09265*. [Online]. Available: <http://arxiv.org/abs/1911.09265>
- [37] J. Yang, N. C. Dvornek, F. Zhang, J. Chapiro, M. Lin, and J. S. Duncan, “Unsupervised domain adaptation via disentangled representations: Application to cross-modality liver segmentation,” in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2019, pp. 255–263.
- [38] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, and M.-H. Yang, “Diverse image-to-image translation via disentangled representations,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 35–51.
- [39] Y. Yang and S. Soatto, “FDA: Fourier domain adaptation for semantic segmentation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4085–4095.
- [40] Y. Yang, D. Lao, G. Sundaramoorthi, and S. Soatto, “Phase consistent ecological domain adaptation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9011–9020.
- [41] N. R. Mollet, S. Dymarkowski, and J. Bogaert, “MRI and CT revealing carcinoid heart disease,” *Eur. Radiol.*, vol. 13, no. S06, pp. L14–L18, Dec. 2003.
- [42] C. Bian *et al.*, “Uncertainty-aware domain alignment for anatomical structure segmentation,” *Med. Image Anal.*, vol. 64, Aug. 2020, Art. no. 101732.
- [43] K. Li, L. Yu, S. Wang, and P.-A. Heng, “Towards cross-modality medical image segmentation with online mutual knowledge distillation,” in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 1, pp. 775–783.
- [44] Q. Dou, Q. Liu, P. Ann Heng, and B. Glocker, “Unpaired multi-modal segmentation via knowledge distillation,” 2020, *arXiv:2001.03111*. [Online]. Available: <http://arxiv.org/abs/2001.03111>
- [45] W. Li, H. Liao, S. Miao, L. Lu, and J. Luo, “Unsupervised learning of landmarks based on inter-intra subject consistencies,” 2020, *arXiv:2004.07936*. [Online]. Available: <http://arxiv.org/abs/2004.07936>
- [46] F. L. Bookstein, “Principal warps: Thin-plate splines and the decomposition of deformations,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, no. 6, pp. 567–585, Jun. 1989.
- [47] T. Jakab, A. Gupta, H. Bilen, and A. Vedaldi, “Self-supervised learning of interpretable keypoints from unlabelled videos,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8787–8797.
- [48] K. M. Yi, E. Trulls, V. Lepetit, and P. Fua, “Lift: Learned invariant feature transform,” in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 467–483.
- [49] Q. Dou, C. Ouyang, C. Chen, H. Chen, and P.-A. Heng, “Unsupervised cross-modality domain adaptation of ConvNets for biomedical image segmentations with adversarial loss,” 2018, *arXiv:1804.10916*. [Online]. Available: <http://arxiv.org/abs/1804.10916>
- [50] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, “The SYNTHIA dataset: A large collection of synthetic images for semantic segmentation of urban scenes,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3234–3243.
- [51] M. P. Heinrich and J. Oster, “MRI whole heart segmentation using discrete nonlinear registration and fast non-local fusion,” in *Proc. Int. Workshop Stat. Atlases Comput. Models Heart.* Cham, Switzerland: Springer, 2017, pp. 233–241.
- [52] C. Ouyang, K. Kamnitsas, C. Biffi, J. Duan, and D. Rueckert, “Data efficient unsupervised domain adaptation for cross-modality image segmentation,” in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2019, pp. 669–677.