2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 978-1-6654-0126-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/BIBM52615.2021.9669615

Tri-Directional Tasks Complementary Learning for Unsupervised Domain Adaptation of Cross-modality Medical Image Semantic Segmentation

Chen Li†, Wei Chen†✉, Mingfei Wu, Xin Luo, Yulin He, Yusong Tan
College of Computer, National University of Defense Technology
Changsha, China
{lichen14, chenwei, mingfeiwu, luoxin13, heyulin, ystan}@nudt.edu.cn

***Abstract*—Cross-modality adaptation is challenging due to the internal domain discrepancy in appearance and representation. When the trained model of source domain is transferred to the target domain, the domain shift will reduce accuracy. Meanwhile, unsupervised domain adaptation has the potential to recover this degradation among medical images of different modalities, so it is of clinical significance and meaningful in bioinformatics. However, previous related works usually try to align domains in a single direction or two directions, failing to take advantage of the complementary relationship between different directions and alignment tasks. In this paper, we propose the Tri-directional learning framework to solve domain shift in the task of medical image semantic segmentation. The proposed framework is able to synergize image style transformation, mask segmentation and edge segmentation. The above three tasks are mutually boosted through complementary training in each iteration. In this way, our method performs cross-modality medical images semantic segmentation from labeled source domain (MRI) to unlabeled target domain (CT). The experimental results demonstrate the effectiveness of the proposed method. For the task of cardiac structure segmentation from cross-modality medical images, our proposed framework achieves state-of-the-art performance. The code is available at https://github.com/lichen14/TriDL.**

***Index Terms*—Deep learning, Unsupervised domain adaptation, Tri-directional Collaborative Learning, Tri-tasks Complementary Boosting, Cardiac structure segmentation.**

I. Introduction

In the research of computer vision (CV), there are obvious barriers to transferring knowledge from one domain to another domain. For example, how to teach a model trained on synthetic images to segment real-world scenes and achieve promising accuracy is extremely challenging. The above phenomenon occurs frequently and exists in many other application scenarios, which results in performance degradation. This is often due to the discrepancy between the cross-modality training and validation. This challenge is called domain shift and has become a hot research issue.

The domain shift also exists in the study of bioinformatics and has an influence on the efficiency of clinical diagnosis. Specifically, in the task of medical image analysis based on deep learning, most of previous works paid attention to the supervised learning, which requires a large number of high-quality annotated samples. The quality and quantity of training dataset directly effect the optimization. As a consequence, these methods need to collect expensive data inevitably, and consume a lot of manpower and time to engage in repetitive annotation work. In clinical practice, due to the higher requirements for the knowledge of labeling experts, the collection of large standardized medical image datasets is more difficult than the collection of natural images (such as ImageNet [1]). Consequently, an annotation-free deep learning method, namely unsupervised domain adaptation (UDA), is introduced to address cross-modality medical image analysis.

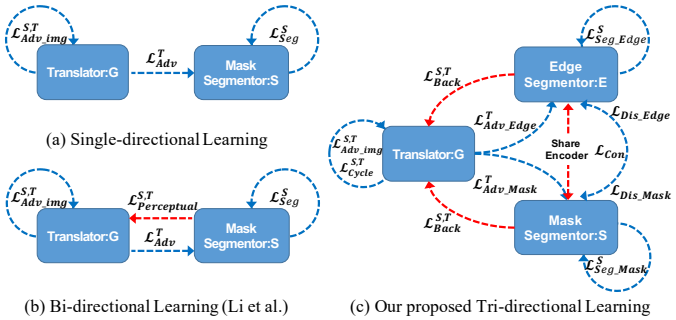

Fig. 1. **Illustration of multi-directional learning.** Single directional learning is shown in Fig.(a) and bi-directional learning is shown in Fig.(b). The Tri-directional learning method we proposed is briefly shown in Fig.(d) and detailed analyzed in Section II.

The main idea of UDA is to extract domain-invariable representations and transfer them from source domain to target domain, where source samples are annotated and the the labels of target samples are absent. Since magnetic resonance imaging (MRI) and computer tomography (CT) are two of the most commonly used methods for clinically scanning abdominal organs, we use both for UDA research in medical image segmentation task, which is illustrated in the Fig.1(a). Specifically, MRI is regarded as the source domain while CT is the target domain. However, the distribution and representation of above two domains are quite different but implicit characteristics are shared. Therefore, it is challenging but practicable to study the cross-modality unsupervised domain adaptation between MRI and CT. And this annotation-free work is of significance to

†These authors contributed equally to this work. This research was funded by the National Key Research and Development Program of China (No. 2018YFB0204301).

bioinformatics and to assist clinicians in diagnosis.

In this paper, we proposed the **Tri-D**irectional **L**earning method to solve domain shift in the task of medical image segmentation, we named it **TriDL**. The TriDL is to carry out collaborative learning among three components, including image translator G, mask segmentor M and edge segmentor E. The above three components perform three tasks respectively, cross-modality style transformation, mask segmentation and edge segmentation. In the iterative training process, these tasks complement and promote each other. In this way, TriDL is trained to perform cross-modality medical images segmentation from labeled source domain to unlabeled target domain. The illustration of multi-directional learning methods is illustrated in the Fig.1. The contributions are as follows:

- We propose the Tri-tasks to work collaborately and extract the general representations for cross-modality adaptation of medical images.
- We propose the Tri-directional complementary learning framework, which boosts translator and segmentors by iteratively training from three directions. To the best of our knowledge, we are the first to propose the Tri-directional learning method to solve unsupervised domain adaptation in bioinformatics.
- We evaluate the methodology on the public cross-modality medical image dataset (MMWHS2017). Our TriDL outperforms other SOTA UDA methods on semantic segmentation of cardiac structures.

II. METHODOLOGIES

In this section, we introduce the novel Tri-directional learning framework (TriDL), which is proposed to solve unsupervised cross-modality domain adaptation in medical image segmentation. We design the Tri-tasks to extract general representations of target domain from labeled source domain (in Section II-A). And the corresponding Tri-directional learning is to coordinate and advance various tasks (in Section II-B).

A. Tri-Tasks Leanring for Unsupervised Domain Adaptation

In this section, we will describe the design of three tasks (i.e., cross-modality style transformation, mask segmentation, edge segmentation) and the network architecture (i.e., G, M, E) in details. The whole framework is shown in the Fig. 2.

The first task is the cross-modality style transformation. Based on the CycleGAN [2], there are two generators (G_{T2S} and G_{S2T}) and discriminators (D_S and D_T) to adversarially generate image with new style. Building on the encoder-decoder architecture, generators extract the implicit features and transform the style of input image. For example, the generator G_{T2S} translates target images x_t into source style-like images x_{t2s} by mapping the representation of target images x_t to source style-like images x_{t2s}. Building on the fully convolutional architecture, discriminators classify the domain of input image. For example, the discriminator D_S distinguishes between real source images x_s and translated fake source images x_{t2s}. The goal of generators is to make synthetic images look similar to real images while discriminators aim to classify all images correctly. Above zero-sum game adversarial learning process can be formulated as the following loss function $\mathcal{L}^S_{Adv\_img}$:

$$\begin{aligned}\mathcal{L}^S_{Adv\_img} = &\textstyle\sum_{x_s\in\mathbb{X}_S}\left[\log D_S(x_s)\right] + \\ &\textstyle\sum_{x_t\in\mathbb{X}_T}\left[\log\left(1-D_S(G_{T2S}(x_t))\right)\right].\end{aligned} \tag{1}$$

Similarity, the adversarial learning function $\mathcal{L}^T_{Adv\_img}$ is defined in the same way to translate source images x_s into target style-like images x_{s2t}.

$$\begin{aligned}\mathcal{L}^T_{Adv\_img} = &\textstyle\sum_{x_t\in\mathbb{X}_T}\left[\log D_T(x_t)\right] + \\ &\textstyle\sum_{x_s\in\mathbb{X}_S}\left[\log\left(1-D_T(G_{S2T}(x_s))\right)\right].\end{aligned} \tag{2}$$

Besides, our work refers the CycleGAN [2] and designs cycle-consistent loss function $\mathcal{L}_{Cycle}$ to avoid contradiction between cross-modality generators when adversarial training, which is formulated as follows:

$$\begin{aligned}\mathcal{L}_{Cycle} = &\textstyle\sum_{x_s\in\mathbb{X}_S}\left[\|G_{T2S}(G_{S2T}(x_s))-x_s\|_1\right] + \\ &\textstyle\sum_{x_t\in\mathbb{X}_T}\left[\|G_{S2T}(G_{T2S}(x_t))-x_t\|_1\right].\end{aligned} \tag{3}$$

The second task is mask segmentation with cross-domain adaptation. Due to this single task is not the main innovation of our work, we follow the success of DeepLab V2 [3] in semantic segmentation task and use it as the mask segmentor. In order to conduct fair comparison with the state-of-the-art methods, we prefer the mainstream segmentation solution that choose the ResNet101 [4] as the segmentor backbone. The mask segmentor mainly consists of two components, i.e., feature extractor and mask generator. The former one extracts domain-shared semantic information from samples of two domains with the help of convolution layers and down-sampling layers. Based on the extracted features, the mask generator recovers the mask from low resolution with the help of symmetric convolution layers and up-sampling layers. The mask segmentor is trained by the supervised loss function $\mathcal{L}_{Seg\_Mask}$ with labeled source samples.

$$\mathcal{L}_{Seg\_Mask} = \textstyle\sum_{x_s\in\mathbb{X}_S, y_s\in\mathbb{Y}_S}[-y_s\log M(x_s)], \tag{4}$$

After that, in order to transfer the domain-shared representation from source domain to target domain, two domain-specific discriminators are introduced to adversarially train the mask segmentor again, where this part will be illustrated later.

The third task is edge segmentation with cross-domain adaptation. This part is similar to the second task, where edge segmentor also consists of feature extractor and edge generator. Meanwhile, edge generator owns the same framework with mask generator. Specifically, the feature extractors of mask segmentor and edge segmentor share the same weights while the generatos own different weights. Before training the edge segmentor, the labels of source domain samples are pre-processed to obtain edge contour. And then edge segmentor is trained by the supervised loss function $\mathcal{L}_{Seg\_Edge}$. Above two domain-specific discriminators are also used to train edge segmentor again.

$$\mathcal{L}_{Seg\_Edge} = \textstyle\sum_{x_s\in\mathbb{X}_S, y_s\in\mathbb{Y}_S}[-\xi(y_s)\log E(x_s)], \tag{5}$$

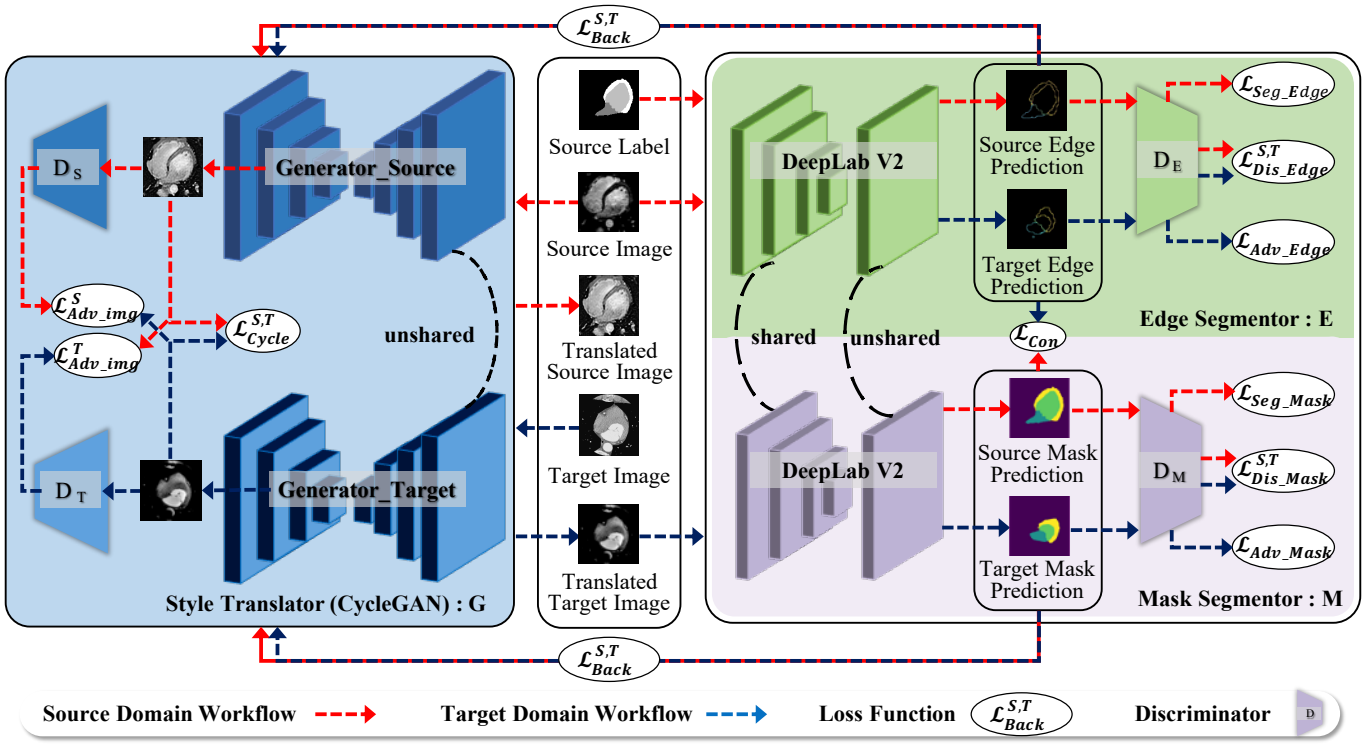

Fig. 2. **An overview of our proposed method.** The whole framework is composed of style translator (G), edge segmentor (E) and mask segmentor (M). The translator is built on the CycleGAN, consisting of two encoder-decoder generators and two discriminators (D_S, D_T). The edge segmentor is built on the DeepLab V2, consisting of features extractor and upsampling layers. Besides, the discriminator D_E is introduced. Similarly, the mask segmentor has the same structure with discriminator D_M. The blue block is the cross-modality style transformation task. The green and pink blocks in the figure respectively represent the edge segmentation and mask segmentation. The cross-modality adaptation is also illustrated by the loss functions. The red arrows and blue arrows represent the data flow of source domain (MRI) and target domain (CT) respectively.

where $\xi(\cdot)$ is the operation of applying the transverse filter and the longitudinal filter to obtain the soft edge of mask from the two spatial axis respectively, which should possess structural consistency with predicted edge. As for the cross-domain adaptation, it is collaborative working with two segmentors to get enhancement, where this part will be illustrated later.

In addition, there are two discriminators D_M and D_E used to distinguish the mask and edge respectively. The discrepancy loss functions are defined as follows:

$$\begin{aligned}\mathcal{L}_{Dis\_Mask} =& \sum_{x_s \in \mathbb{X}_S} \mathcal{L}_{bce}\{D_M[M(x_s)], M\} + \\ & \sum_{x_t \in \mathbb{X}_T} \mathcal{L}_{bce}\{D_M[M(G_{T2S}(x_t))], M\},\end{aligned} \tag{6}$$

$$\begin{aligned}\mathcal{L}_{Dis\_Edge} =& \sum_{x_s \in \mathbb{X}_S} \mathcal{L}_{bce}\{D_E[E(x_s)], E\} + \\ & \sum_{x_t \in \mathbb{X}_T} \mathcal{L}_{bce}\{D_E[E(G_{T2S}(x_t))], E\}.\end{aligned} \tag{7}$$

B. Tri-directional Collaborative Learning

Our proposed complementary learning framework consists of three directions shown in Fig.1(d).

The first direction is the translator boosts the mask segmentor and edge segmentor ($G \rightarrow M$, $G \rightarrow E$). In this direction, the cross-modality translator G is firstly trained with unpaired source domain samples $\mathbb{X}_S$ and target domain sampels $\mathbb{X}_T$. Generative adversarial learning method is utilized to transform sample's style with adversarial loss $\mathcal{L}^S_{Adv\_img}$ (1), $\mathcal{L}^T_{Adv\_img}$ (2) and cycle-consistent loss $\mathcal{L}_{Cycle}$ (3). The optimization of translator G can be formulated with loss functions:

$$\min_{\theta_{G_{S2T}}, G_{T2S}} \max_{\theta_{D_S}, D_T} [\mathcal{L}^S_{Adv\_img} + \mathcal{L}^T_{Adv\_img} + \mathcal{L}_{Cycle}]. \tag{8}$$

After image style transformation, we utilize the labeled source samples $\{\mathbb{X}_S, \mathbb{Y}_S\}$ and unlabeled translated target samples $\mathbb{X}_{T2S}$ to train the segmentors M and E. These segmentors are trained by annotated source samples at first with supervised loss functions $\mathcal{L}_{Seg\_Mask}$ (4) and $\mathcal{L}_{Seg\_Edge}$ (5) respectively. After that, segmentors are adversarially optimized with unlabeled target samples to fool the discriminators:

$$\mathcal{L}_{Adv\_Mask} = \sum_{x_t \in \mathbb{X}_T} \mathcal{L}_{bce}\{D_M[M(G_{T2S}(x_t))], S\}, \tag{9}$$

$$\mathcal{L}_{Adv\_Edge} = \sum_{x_t \in \mathbb{X}_T} \mathcal{L}_{bce}\{D_E[E(G_{T2S}(x_t))], S\}. \tag{10}$$

In summary, we can derive the optimization problem to boost M and E with the help of transformation G:

$$\min_{\theta_M} \max_{\theta_{D_M}} [\mathcal{L}_{Seg\_Mask} + \lambda_{Adv} \mathcal{L}_{Adv\_Mask}], \tag{11}$$

$$\min_{\theta_E} \max_{\theta_{D_E}} [\mathcal{L}_{Seg\_Edge} + \lambda_{Adv} \mathcal{L}_{Adv\_Edge}]. \tag{12}$$

The second direction is the mask segmentor and edge segmentor work collaboratively with each other ($M \leftrightarrow E$). We can obtain well-trained segmentors M and E after the first direction learning. However, the consistency between

mask and edge is not utilized. Therefore, in order to leverage invariable representation in the boundary of the mask, we propose the edge alignment method to improve the performance of segmentors. In this direction, the low-level features, which are extracted by encoder and describe the semantic information of boundary, are shared by the feature extractor of two segmentors M and E. Besides, in order to obey the fact that the prediction of edge and the boundary of mask should keep consistent, we introduce the loss functions ($\mathcal{L}_{Con}$) to align predicted masks and edges in the self-supervised manner.

$$\mathcal{L}_{Con} = \sum_{x_t \in \mathbb{X}_T} \mathcal{L}_{bce} \left\{ \xi(M[G_{T2S}(x_t)]), E[G_{T2S}(x_t)] \right\}, \quad (13)$$

where $\mathcal{L}_{bce}$ is the Binary Cross Entropy loss function, $\xi(\cdot)$ is the edge detection operation in (5). In summary, we can achieve optimization through self-supervised collaboration between M and E:

$$\min_{\theta_{M,E}} \max_{\theta_{D_M, D_E}} \left[\mathcal{L}_{Dis\_Mask} + \mathcal{L}_{Dis\_Edge} + \lambda_{Con} \mathcal{L}_{Con} \right]. \quad (14)$$

The third direction is the well-trained mask segmentor and edge segmentor promote the translator in return ($M \rightarrow G$, $E \rightarrow G$). In this direction, we believe that the best image translation model should preserve the implicit semantic information while only transforming the style of the sample. Based on this motivation, $\mathcal{L}_{Cycle}$ and $\mathcal{L}_{Back}$ are utilized for enhancing translator G. The $\mathcal{L}_{Cycle}$ is already implemented in the (3) at the beginning, which reconstructs the translated image again to optimize style transformation. The $\mathcal{L}_{Back}$ is the backward propagation, which aims to maintain the semantic consistency between original sample and translated sample. For example, the predicted results of source image x_s and translated image x_{s2t} should keep same. The predicted results of source image x_s and reconstructed image x_{s2t2s} should keep same. The loss function of source domain backward optimization $\mathcal{L}^S_{Back}$ is defined as follows. Similarity, the target domain backward optimization $\mathcal{L}^T_{Back}$ is defined in the symmetric way.

$$\begin{aligned} \mathcal{L}^S_{Back} = & \textstyle\sum_{x_s \in \mathbb{X}_S} [\|M(G_{S2T}(x_s)) - M(x_s)\|_1 \\ & + \|M(G_{T2S}(G_{S2T}(x_s))) - M(x_s)\|_1 \\ & + \|E(G_{S2T}(x_s)) - E(x_s)\|_1 \\ & + \|E(G_{T2S}(G_{S2T}(x_s))) - E(x_s)\|_1]. \end{aligned} \quad (15)$$

$$\begin{aligned} \mathcal{L}^T_{Back} = & \textstyle\sum_{x_t \in \mathbb{X}_T} [\|M(G_{T2S}(x_t)) - M(x_t)\|_1 \\ & + \|M(G_{S2T}(G_{T2S}(x_t))) - M(x_t)\|_1 \\ & + \|E(G_{T2S}(x_t)) - E(x_t)\|_1 \\ & + \|E(G_{S2T}(G_{T2S}(x_t))) - E(x_t)\|_1]. \end{aligned} \quad (16)$$

In summary, the third directional learning method connects the well-trained segmentors and translator instead of only optimizing single component. Meanwhile, the semantic information remains consistent during the training process. With the collaboration of segmentors M and E, the optimization of translator is updated , which is defined as:

$$\begin{aligned} \min_{\theta_{G_{S2T}, G_{T2S}}} \max_{\theta_{D_S, D_T}} & [\mathcal{L}^S_{Adv\_img} + \mathcal{L}^T_{Adv\_img} + \mathcal{L}_{Cycle} \\ & + \lambda_{Back}(\mathcal{L}^S_{Back} + \mathcal{L}^T_{Back})]. \end{aligned} \quad (17)$$

C. Network Configurations and Implementation Details

Deep learning environment: We built the Python with version 3.6.9 and PyTorch with version 1.3.0 on the Ubuntu system (18.04) with one NVIDIA 1080Ti GPU.

Cross-modality style transformation task: As shown in the blue block in Fig.2, the CycleGAN [2] is designed as backbone. The inputs of translator G are source image and target image. Generators of source and target are domain-specific, and they have the same encoder-decoder architecture but do not share weights, where encoder is used for features extraction and decoder for reconstruction. Discriminator D_S and D_T also have the same architecture but different weights, which stacked by four layers (Conv, InstanceNorm, LeakyReLU), and finally attach FCN to get the classification result. Networks are trained by the Adam optimizer [5] with the initial learning rate 0.001. And the training batchsize is set to 1. This task is optimized by the Equation (8) and (17).

Mask segmentation task: As shown in the pink block in Fig.2, the DeepLab-v2 [3] is designed as backbone of mask segmentor M, which is built on the auto-encoder structure and consists of encoder for features extraction and upsampling layers for super-resolution. The backbone is pre-trained on the ImageNet [1]. The discriminator D_M is stacked by four layers (Conv, LeakyReLU), and finally attach FCN to get the classification result. This task is optimized by the Equation (11).

Edge segmentation task: As shown in the green block in Fig.2, the structure of edge segmentor E is same as M. Due to the fact that the mask and edge is consistent in the boundary, the extracted features of M and E is similar. Consequently, the encoder of above two segmentors share the weights but upsampling layer is different. This task is optimized by the Equation (12). Both mask segmentation and edge segmentation networks are trained by Adam optimizer [5] with the initial learning rate 0.0001. And the training batchsize is set to 8.

III. Experiment

A. Experimental settings

Experimental dataset: The proposed TriDL is evaluted in the cardiac structures segmentation. The experimental dataset is from the public MICCAI challenge, i.e., Multimodality Whole Heart Segmentation Challenge (MMWHS2017) [18]. The contents of heart are described in different modalities, including 20 CT volumes and 20 MRI volumes. Inside the heart, the labels of seven structures are obtained through pixel-level annotation. As for the dataset division, we follow the previous research [8], [14], [15], [17] in this field, 16 MRI volumes and 16 CT volumes are randomly selected as the training set of the source domain and the target domain respectively, and the remaining 4 MRI volumes and 4 CT volumes are used as the testing set. The labels of testing set are only used for evaluation.

Preprocessing methods: As for label selection, we follow the previous research and choose four cardiac structures for fair comparison, including ascending aorta (AA), left atrium

TABLE I

EXPERIMENTAL RESULTS COMPARISON WITH SUPERVISED LEARNING AND UNSUPERVISED DOMAIN ADAPTATION FOR 4 CARDIAC ORGANS SEGMENTATION. THE EVALUATION METRIC IN THIS TABLE IS DICE (MEAN±STD). [1]

Methods (Dice)	Ascending Aorta (AA)	Left Atrium blood Cavity (LAC)	Left Ventricle blood Cavity (LVC)	Myocardium of left ventricle (MYO)	Average
Without adapation	27.57±16.37	27.63±14.29	33.79±8.81	13.81±9.44	25.70±0.78
ADDA [6]	47.60	60.90	11.20	29.20	37.20
DANN [7]	39.00	45.10	28.30	25.70	34.50
Pnp-AdaNet [8]	74.00±7.30	68.90±5.20	61.90±10.70	50.80±7.00	63.90±7.50
SynSeg-Net [9]	71.60	69.00	51.60	40.80	58.20
CycleGAN [2]	73.80	75.70	52.30	28.70	57.60
CyCADA [10]	72.90	77.00	62.40	45.30	64.40
BEAL [11]	75.47±9.67	62.77±9.47	68.49±11.23	57.93±7.59	66.17±10.25
Cascaded U-Net [12]	77.36±6.28	65.28±10.16	70.05±9.60	60.66±9.74	68.34±9.31
SIFA-v1 [13]	81.10	76.40	75.70	58.70	73.00
SIFA-v2 [14]	81.30	79.50	73.80	61.60	74.10
DualHierNet [15]	84.70±6.41	74.61±10.01	**83.42±7.46**	**65.19±6.33**	76.98±7.84
BDL [16]	84.34±6.76	71.31±18.70	77.04±3.50	60.36±13.73	73.26±10.31
DSFN [17]	84.70	76.90	79.10	62.40	75.80
TriDL (ours)	**88.87±6.70**	**78.85±15.64**	78.64±2.70	62.27±9.69	**77.16±6.70**
Supervised learning	92.43±2.31	83.84±8.87	91.09±4.72	85.39±6.75	88.18±4.47

blood cavity (LA), left ventricle blood cavity (LV) and myocardium of the left ventricle (MYO). In order to better observe and diagnose heart, the original 3D volume were firstly manually cropped to cover the any of the above cardiac structures, and then split into transverse view slices as 2D inputs. All the inputs were normalized as zero mean and unit variance with size of 256×256.

Evaluation metrics: The Dice similarity coefficient (Dice) [19] was widely used in previous cross-modality domain adaptation works [8], [13], [14], [17] to quantitatively compare the performance discrepancy of methods. Dice is used to calculate the overlap between predicted result and ground truth. Higher Dice values represent better performance.

B. Comparison with the State-of-the-art Methods

Before the comparison, we firstly measured the domain shift by conducted experiments to get the performance gap in the cardiac structures segmentation. The bottom bound was to directly transfer the well-trained model from source domain to target domain without any domain adaptation technique. The top bound was to use the labels of target samples to train the target model based on supervised learning. The numerical difference in performance values between the bottom bound and top bound was regarded as the domain shift, which were reported in the Table I. The bold number highlighted the best performance.

We found that the model trained on MRI can only get an average Dice of 25.70±0.78% when it was directly used to segment CT images without any domain adaptation. But when the supervised learning model was trained from CT samples and then used to segment CT images, an average Dice of 88.18±4.47% is obtained. As a consequence, we can obtain the huge performance gap by calculating the discrepancy between top bound and bottom bound, which also indicated the severe domain shift between cross-modality images (MRI → CT) and the challenge of domain-adaptive medical image segmentation.

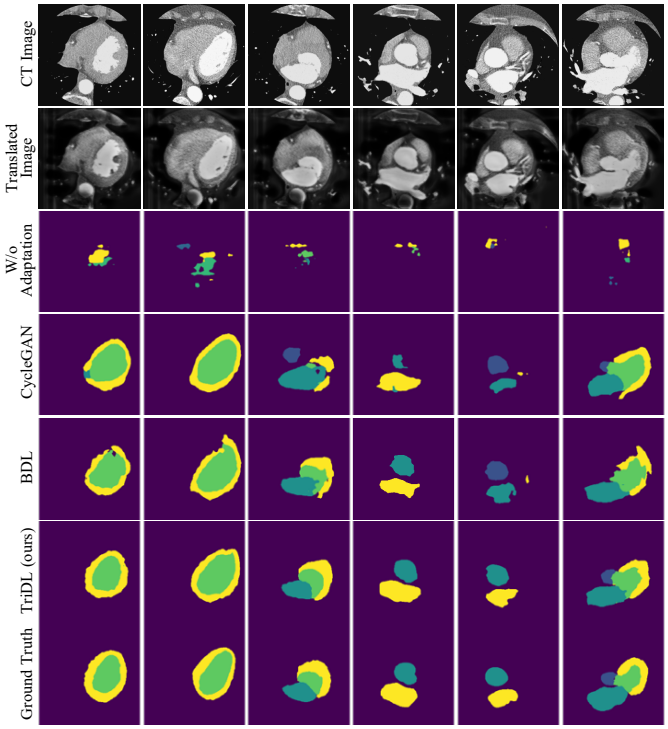

Fig. 3. **Visual comparison of segmentation results predicted by different methods for cardiac structures.** The top row shows seven original CT images, and each column corresponds to a sample. Corresponding from top to bottom are translated images (2nd row), the predictions without adaptation (3rd row), the predictions of CycleGAN (4th row), the predictions of SIFA (5th row), the predictions of BDL (6th row), the predictions of our proposed TriDL (7th row), the predictions of supervised training (8th row), and ground truth (last row). Different cardiac structures (AA, LAC, LVC, MYO) are represented in different colors.

In order to fairly evaluate the discrepancy between the proposed TriDL and other UDA methods, we applied other twelve models to segment cardiac structures from cross-modality medical images. The quantitative evaluated results

were reported in the Table I, where the results of above SOTA methods all referred from their papers. Specifically, the results of [8] come from the Table 3 in its paper, the results of [17] come from the Table 2 in its paper, the results of [15] come from the Table 2 in its paper, the results of [13] come from the Table 1 in its paper, the results of [14] come from the Table 1 in its paper. In addition, since there were no relevant experimental data in the original papers, the results of [6], [7] come from the Table 1 of paper [13], the results of [2], [9], [10] come from the Table 1 of paper [14], the results of [11], [12] come from the Table 2 of paper [15].

In the segmentation of cardiac structures, our proposed TriDL achieved a significant improvement by achieving an average Dice of $77.16 \pm 6.70\%$ on four cardiac structures. On the one hand, this results were very competitive and overall better than other SOTA methods. For example, compared with the SOTA method SIFA-v2 [14], our TriDL increased the average Dice from 74.10% to 77.16%. Compared with another SOTA method BDL [16], our TriDL increased the average Dice from 74.10% to 77.16%. Overall, TriDL owned the smallest standard deviation but highest mean, which indicated that our method was stable and performed well. On the other hand, this results were very close to the results of supervised training and also proved the effectiveness of the proposed TriDL.

We also reimplemented above methods under the same settings for qualitative analysis. We visually compared the segmentation results and displayed the five groups of predictions, which were shown in the Fig.3. We found that it was difficult to obtain correct contour for any cardiac structure without adaptation technique. Instead, the results of supervised training were similar to the ground truth. The discrepancy of above two results demonstrated the server domain shift in the cross-modality medical image segmentation. Compared with other SOTA methods, our method can successfully locate the four structures and generate semantically meaningful mask.

Both the quantitative results and qualitative results of cross-modality medical image segmentation validated the effectiveness of our Tri-directional tasks complementary learning method for unsupervised domain adaptation.

IV. Conclusion

In this paper, we propose an unsupervised domain-adaptive framework (TriDL) for cross-modality medical image semantic segmentation. TriDL synergizes three tasks for domain generalization, including cross-modality style transformation and mask segmentation and edge segmentation. These tasks collaborately work to learn the domain-adaptive representations and effectively reduce the performance degradation during cross-domain adaptation without any target annotations. Meanwhile, through the Tri-directional collaborative learning method, the translator and the segmentors promote each other. Experiments show that our work reaches the state-of-the-art in the domain adaptive segmentation of cross-modality medical images on the heart structures dataset.

References

[1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 248–255.

[2] J. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2242–2251.

[3] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 40, no. 4, pp. 834–848, 2018.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[5] S. Bock and M. Weiß, "A proof of local convergence for the adam optimizer," in *2019 International Joint Conference on Neural Networks (IJCNN)*, 2019, pp. 1–8.

[6] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2962–2971.

[7] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. March, and V. Lempitsky, "Domain-adversarial training of neural networks," *Journal of Machine Learning Research*, vol. 17, no. 59, pp. 1–35, 2016.

[8] Q. Dou, C. Ouyang, C. Chen, H. Chen, B. Glocker, X. Zhuang, and P. Heng, "Pnp-adanet: Plug-and-play adversarial domain adaptation network at unpaired cross-modality cardiac segmentation," *IEEE Access*, vol. 7, pp. 99 065–99 076, 2019.

[9] Y. Huo, Z. Xu, H. Moon, S. Bao, A. Assad, T. K. Moyo, M. R. Savona, R. G. Abramson, and B. A. Landman, "Synseg-net: Synthetic segmentation without target modality ground truth," *IEEE Transactions on Medical Imaging (TMI)*, vol. 38, no. 4, pp. 1016–1025, 2019.

[10] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. Efros, and T. Darrell, "CyCADA: Cycle-consistent adversarial domain adaptation," in *Proceedings of the 35th International Conference on Machine Learning (ICML)*, vol. 80, 2018, pp. 1989–1998.

[11] S. Wang, L. Yu, K. Li, X. Yang, C.-W. Fu, and P.-A. Heng, "Boundary and entropy-driven adversarial learning for fundus image segmentation," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, 2019, pp. 102–110.

[12] C. Chen, C. Ouyang, G. Tarroni, J. Schlemper, H. Qiu, W. Bai, and D. Rueckert, "Unsupervised multi-modal style transfer for cardiac mr segmentation," in *Statistical Atlases and Computational Models of the Heart. Multi-Sequence CMR Segmentation, CRT-EPiggy and LV Full Quantification Challenges*, 2020, pp. 209–219.

[13] 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," *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, vol. 33, no. 01, pp. 865–872, 2019.

[14] 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 Transactions on Medical Imaging (TMI)*, vol. 39, no. 7, pp. 2494–2505, 2020.

[15] Y. Xue, S. Feng, Y. Zhang, X. Zhang, and Y. Wang, "Dual-task self-supervision for cross-modality domain adaptation," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*. Cham: Springer International Publishing, 2020, pp. 408–417.

[16] Y. Li, L. Yuan, and N. Vasconcelos, "Bidirectional learning for domain adaptation of semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[17] D. Zou, Q. Zhu, and P. Yan, "Unsupervised domain adaptation with dual-scheme fusion network for medical image segmentation," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI)*, 7 2020, pp. 3291–3298.

[18] X. Zhuang and J. Shen, "Multi-scale patch and multi-modality atlases for whole heart segmentation of mri," *Medical Image Analysis*, vol. 31, pp. 77–87, 2016.

[19] C. Li, Y. Tan, W. Chen, X. Luo, Y. He, Y. Gao, and F. Li, "Anu-net: Attention-based nested u-net to exploit full resolution features for medical image segmentation," *Computers & Graphics*, 2020.