Unsupervised Domain Adaptation with Dual-Scheme Fusion Network for Medical Image Segmentation

Danbing Zou[1] , **Qikui Zhu**[1*] and **Pingkun Yan**[2*]
[1]School of Computer Science, Institute of Artificial Intelligence, and National Engineering Research Center for Multimedia Software, Wuhan University, China
[2]Rensselaer Polytechnic Institute, Department of Biomedical Engineering, Troy, NY, USA
{danbingzou, QikuiZhu}@whu.edu.cn, yanp2@rpi.edu

Abstract

Domain adaptation aims to alleviate the problem of retraining a pre-trained model when applying it to a different domain, which requires large amount of additional training data of the target domain. Such an objective is usually achieved by establishing connections between the source domain labels and target domain data. However, this imbalanced source-to-target one way pass may not eliminate the domain gap, which limits the performance of the pre-trained model. In this paper, we propose an innovative Dual-Scheme Fusion Network (DSFN) for unsupervised domain adaptation. By building both source-to-target and target-to-source connections, this balanced joint information flow helps reduce the domain gap to further improve the network performance. The mechanism is further applied to the inference stage, where both the original input target image and the generated source images are segmented with the proposed joint network. The results are fused to obtain more robust segmentation. Extensive experiments of unsupervised cross-modality medical image segmentation are conducted on two tasks – brain tumor segmentation and cardiac structures segmentation. The experimental results show that our method achieved significant performance improvement over other state-of-the-art domain adaptation methods.

1 Introduction

Deep learning has achieved state-of-the-art results in many fields. However, the model trained on one dataset cannot be directly applied to other datasets with similar contents and various distribution. Since the different distribution brings a domain shift problem and further leads to identification failure. Solving the identification failure caused by domain shift is an important issue, which can improve the generalization as well as portability and further solve the problem of lacking of labeled data and expensive cost of manual annotation. Unsupervised domain adaptation methods are an effective method to solve those problems, which enables one labeled dataset to serve multiple other unlabeled ones and greatly improve the practical value of deep learning network. The unsupervised domain adaptation methods can be categorized into two types: image adaptation and feature adaptation. [Chen *et al.*, 2018; Bousmalis *et al.*, 2017; Hoffman *et al.*, 2017; Chen *et al.*, 2019] are the representative articles of image adaptation. In [Chen *et al.*, 2018; Bousmalis *et al.*, 2017], segmentation network is first trained using source domain dataset, and then target data are translated into source domain for segmentation using the trained segmentation network. In [Hoffman *et al.*, 2017; Chen *et al.*, 2019], the source domain data are first translated into target domain, and then the segmentation network are trained by the generated target-like data with corresponding source domain annotation. The representative articles of feature adaptation are [Tsai *et al.*, 2018; Dou *et al.*, 2018]. Those methods directly perform feature adaptation without cross-domain translation of the image.

However, current works based on image adaptation only conduct the two-direction translation separately. This imbalanced source-to-target or target-to-source one way connection may not eliminate the domain gap, which limits the performance of the pre-trained model. To overcome this problem, we propose to combine these two complementary directions into one unified framework for acquiring a better result. Our proposed model, named Dual-Scheme Fusion Network (DSFN), adopts CycleGAN[Zhu *et al.*, 2017]'s architecture as base structure. To make the network can simultaneously conduct the tasks of image translation and segmentation, we modified the generator to a joint translation-segmentation network. The translation and segmentation parts share a same encoder. This setting can not only reduce network's parameters, but also enable the two tasks to enhance each other. Moreover, the segmentation network also can help reserve content of the generated image. During training, the first joint cross-domain translation and segmentation network is trained by source domain images. The task of translation branch is generating target-like images and the segmentation branch aims to segment the source domain image. The first trained segmentation network can be used to segment the source-like image generated by the target domain image. Then, the generated target-like image is used to train the joint network for converting target-like image to source domain and segmen-

*Corresponding Authors.

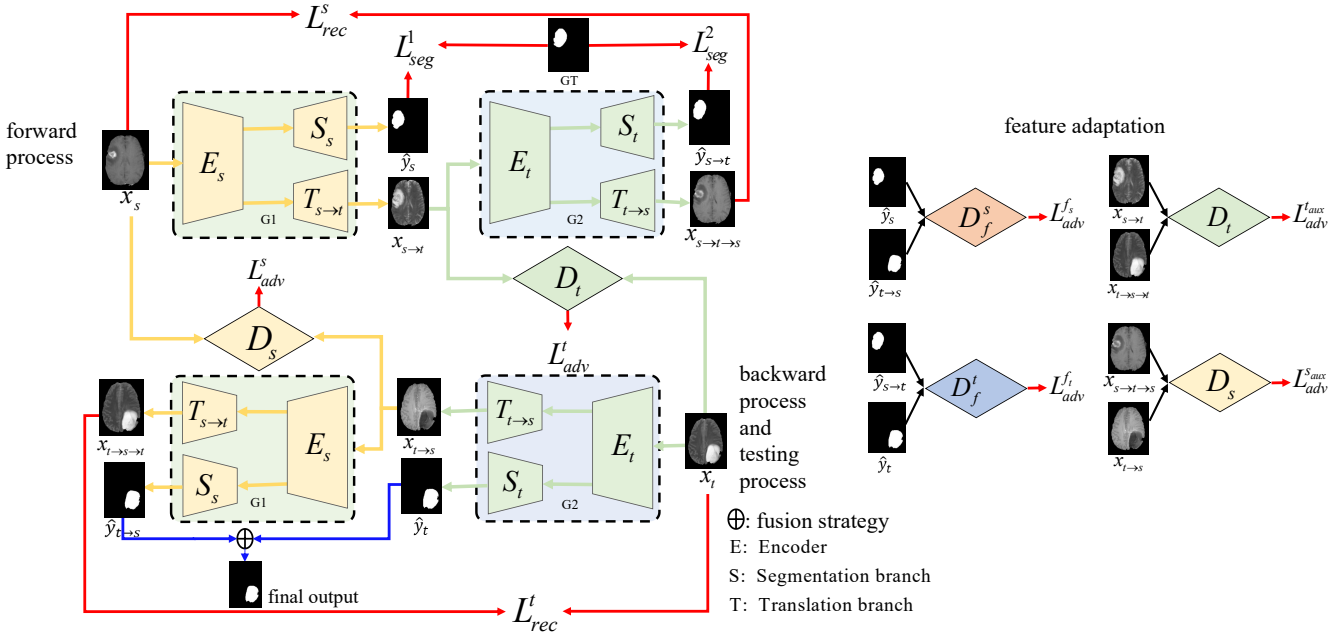


Figure 1: Overall structure of our proposed network. $E_s + T_{s\to t}$ is the source-to-target image translation branch while $E_t + T_{t\to s}$ conduct target-to-source translation. $E_s + S_s$ and $E_t + S_t$ are the two segmentation network that we aim to get. We add two other discriminators D_f^s, D_f^t and auxiliary tasks for D_t, D_s to perform feature adaptation. The red arrows represent different losses, the yellow arrows show the data flows from source or source-like images, the green arrows show the data flows from target or target-like images, and the black arrows mean the data flows for feature adaptation.

tation. The second obtained segmentation branch can be directly used to segment the target domain data. Finally, we merge the two segmentation outputs of target image to obtain a more accurate result. Moreover, we also introduce feature adaptation loss referring[Chen *et al.*, 2019] to further narrow the feature-level gap between target and target-like data, as well as source and source-like data. The whole network is designed subtly, the backward process of our network is also the test procedure. The main contributions of our paper are as follows:

1) We propose an end-to-end unsupervised parallel dual-scheme fusion network for solving the problem of unsupervised domain adaptation from two perspectives by building both source-to-target and target-to-source connections, this balanced joint information flow helps bridge the domain gap to further improve the network performance.
2) We train our network in a cycle manner and the backward process is also the testing process. The structure of our network is effective and elegant.
3) We validate our method on the BRATS[Menze *et al.*, 2014] and MM-WHS[Zhuang and Shen, 2016] dataset. Experimental results show that our proposed approach exceed current state-of-the-art domain adaptation methods, effectively alleviating the expensive cost of manual labeling cross-domain medical data, which has great clinical practical value.

2 Related Work

Typically, domain adaptation methods generalize to other tasks by minimizing the distance of feature space between the source and the target domain. For example, Tzeng et al. [Tzeng *et al.*, 2014] tries to learn a domain-invariant feature representation between source and target by proposing an adaptation layer and a domain confusing loss. Long et al. [Long *et al.*, 2015] matches source and target distribution by embedding their hidden features in a reproducing kernel Hilbert space.

Inspired by the advanced of generative adversarial network [Goodfellow *et al.*, 2014], adversarial learning becomes the main training strategy for domain adaptation methods. Some of them directly map feature representations between source and target domain, for example, Tsai et al .[Tsai *et al.*, 2018] conducted different feature levels' domain adaptation by introducing multi-level discriminators to the segmentation network and Dou et al. [Dou *et al.*, 2018] shared a similar idea with an innovative network structure. Some conduct pixel-level domain transfer, which has two directions as mentioned above. One is to perform target-to-sourced translation, such as [Chen *et al.*, 2018], and the other is to perform source-to-

Method	Dice Score [%]				Hausdorff Distance [mm]			
	T1	FLAIR	T1CE	Average	T1	FLAIR	T1CE	Average
Without domain adaptaion	6.8	54.4	6.7	22.6	58.7	21.5	60.2	46.8
CycleGAN[Zhu *et al.*, 2017]	38.1	63.3	42.1	47.8	25.4	17.2	23.2	21.9
CyCADA [Hoffman *et al.*, 2017]	49.6	72	51.7	57.8	20.2	14.9	18.4	17.8
AdaptSegNet[Tsai *et al.*, 2018]	42.6	67.8	33.1	47.8	23.4	16.6	28.8	22.9
SIFA[Chen *et al.*, 2019]	51.7	68	58.2	59.3	19.6	16.9	**15.01**	17.1
DSFN(our proposed)	**57.3**	**78.9**	**62.2**	**66.1**	**17.5**	**13.8**	15.5	**15.6**

Table 1: Comparison of Dice score and Hausdorff distance between our proposed method and other state-of-the-art methods on whole tumor segmentation in the BRATS dataset.

target translation, such as [Hoffman *et al.*, 2017] and [Zhu *et al.*, 2020].

Due to the expensive cost of manual annotation, there is a lot of research on domain adaptation in the medical field. For example, Nie et al. [Nie *et al.*, 2017] proposed a context-aware GAN to generate PET images from MRI images. Wolterink et al. [Wolterink *et al.*, 2017] used CycleGAN to convert between unpaired CT and MRI images. Similarly, Chen et al. [Chen *et al.*, 2019] conducted the MRI to CT translation and CT splenomegalia segmentation at the same time. Zhao et al. [Zhao *et al.*, 2018] proposed a R-sGAN technique to synthesize fundus image data set and generalize its segmentation network to other data sets. Chen et al. [Chen *et al.*, 2018] adopted target-to-source translation to segment left/right lung using chest X-ray datasets. Kamnitsas et al. [Kamnitsas *et al.*, 2017] proposed a multi-connected domain discriminator to segment unannotated MR images with traumatic brain injuries. [Dou *et al.*, 2018] transfers a cardiac segmentation network trained with MRI to unannotated CT images. Fang and Yan [Fang and Yan, 2020] propose a new multi-scale paramid network and loss function to segment multiple partially labeled datasets.

3 Methods

The overall framework of our proposed Dual-Scheme Fusion Network for unsupervised domain adaptation is shown in Fig.1, which contains two joint translation-segmentation modules. Each joint network has an encoder-decoder structure. The decoder is divided into two branches–translation branch and segmentation branch, which share one same encoder. Each translation branch corresponds to a discriminator to distinguish the generated image from the real image. The design of our network structure is a transformation based on CycleGAN[Zhu *et al.*, 2017]. The difference is that we replace the generator with a joint translation-segmentation module.

The background setting is that we have a source domain dataset $\{x_s\}$ with corresponding annotation $\{y_s\}$ and want to annotate the unlabeled target domain dataset $\{x_t\}$. We solve this problem from two perspectives: (a) Target-to-Source: We train a segmentation network E_s + S_s using source domain dataset $\{x_s, y_s\}$, then translate target image x_t into the source domain through E_t + $T_{t\to s}$ and segment it through the segmentation network E_s + S_s; (b) Source-to-Target: we first translate the source domain image x_s into the target domain through E_s + $T_{s\to t}$, then a segmentation network E_t + S_t is trained using the generated $x_{s\to t}$ with corresponding annotation y_s, so the real x_t can be directly segmented by E_t + S_t. We integrate these two streams into a unified network and fuse their results to form a new prediction as final output.

The flow of data through our network is described as follows. In the training process of forward, x_s is inputed into the joint network G_1 to obtain the cross-domain generated image $x_{s\to t}$ and the segmentation prediction $\hat{y}_s$. Discriminator D_t will distinguish between x_t and $x_{s\to t}$. The generated image $x_{s\to t}$ is then putted into the joint network G_2 to obtain reconstructed image $x_{s\to t\to s}$ and the segmentation prediction $\hat{y}_{s\to t}$.

And in the training process of backward (test process follows the same data flow of backward process), the target domain image x_t is inputted into joint network G_2, and we can get the x_s-like cross-modality image $x_{t\to s}$ and segmentation prediction $\hat{y}_t$. Discriminator D_s distinguish between x_s and $x_{t\to s}$. Then, the generated $x_{t\to s}$ is putted into the joint network G_1 which outputs the reconstructed image $x_{t\to s\to t}$ and the segmentation prediction $\hat{y}_{t\to s}$. As for feature adaptation, Discriminator D_f^s distinguishes between $\hat{y}_s$ and $\hat{y}_{t\to s}$ while D_f^t will distinguish between $\hat{y}_{s\to t}$ and $\hat{y}_t$. D_t and D_s also have an auxiliary task to respectively distinguish between $x_{s\to t}$ and $x_{t\to s\to t}$, $x_{s\to t\to s}$ and $x_{t\to s}$. In the test process, $\hat{y}_t$ and $\hat{y}_{t\to s}$ are two segmentation predictions for the target domain image x_t from different perspective. We fuse the two predictions to get a higher-accuracy result fusion($\hat{y}_t$, $\hat{y}_{t\to s}$) as our final result for target image x_t. For the selection of the fusion strategy, we find that averaging the prediction probabilities of these two results can obtain better performance.

3.1 Adversarial Loss

We train our unified network in a cycle manner, which means that we have a source to target translation and a target to source translation as well. To generate more realistic images, we adopt the adversarial loss as loss function. The adversarial learning consists of two modules: a generator and a discriminator. Our problem is an image-to-image domain translation problem. So in source to target translation, we forward a source image x_s to generator E_s+$T_{s\to t}$ and it's responsible for generating an image $x_{s\to t}$ that looks like target image and can fool the discriminator D_t, while the discriminator D_t aims to distinguish whether the image is actually derived from target domain or synthesized by the generator. This forms a

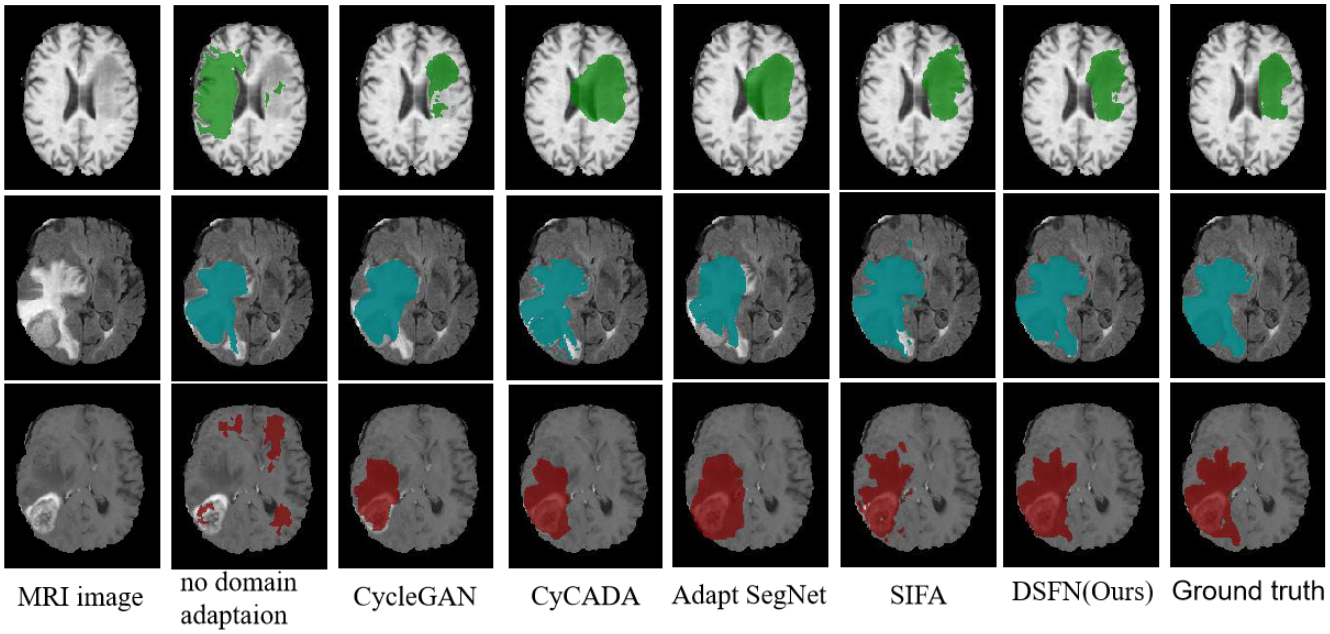


Figure 2: Visual display of the results for comparing different methods on BRATS dataset. From left to right are the target domain input images, results without domain adaptation (as the baseline), CycleGAN, CYCADA, Adapt Segnet, SIFA and our proposed method. The last column is ground truth. The first row is T1 images with segmentation in green, the second row is FLAIR images with segmentation in blue, and the last row is T1CE images with segmentation in red.

process of adversarial learning, which can be described as $min_{E_s+T_{s\to t}} max_{D_t} L_{adv}(E_s, T_{s\to t}, D_t)$ and:

$$L_{adv}(E_s, T_{s\to t}, D_t) = E_{x_t\sim X_t}\left[logD_t\left(x_t\right)\right] + E_{x_s\sim X_s}\left[log\left(1 - D_t\left(T_{s\to t}\left(E_s\left(x_s\right)\right)\right)\right)\right] \tag{1}$$

The target to source translation is likewise. The objective function is $min_{E_t+T_{t\to s}} max_{D_s} L_{adv}(E_t, T_{t\to s}, D_s)$ and:

$$L_{adv}(E_t, T_{t\to s}, D_s) = E_{x_s\sim X_s}\left[logD_s\left(x_s\right)\right] + E_{x_t\sim X_t}\left[log\left(1 - D_s\left(T_{t\to s}\left(E_t\left(x_t\right)\right)\right)\right)\right] \tag{2}$$

3.2 Reconstruction Loss

In order to make the generated image preserve structure, shape and other content information of the original image, we introduce reconstruction loss in our model to constraint the training process.

Taking the forward process as example, reconstruction means that $x_{s\to t}$ translated by x_s can be reconstructed back to source domain by generator E_t+$T_{t\to s}$ and will be consistent with the original x_s, ie $T_{t\to s}\left(E_t\left(x_{s\to t}\right)\right) \sim x_s$. Its objective function is:

$$L_{rec}\left(E_s, T_{s\to t}, E_t, T_{t\to s}\right) = E_{x_s\sim X_S}\left[\left\|T_{t\to s}\left(E_t\left(\left(T_{s\to t}\left(E_s\left(x_s\right)\right)\right)\right)\right) - x_s\right\|_1\right] \tag{3}$$

Similarly, in the backward process, it is $T_{s\to t}\left(E_s\left(x_{t\to s}\right)\right) \sim x_t$:

$$L_{rec}\left(E_s, T_{s\to t}, E_t, T_{t\to s}\right) = E_{x_t\sim X_T}\left[\left\|T_{s\to t}\left(E_s\left(\left(T_{t\to s}\left(E_t\left(x_t\right)\right)\right)\right)\right) - x_t\right\|_1\right] \tag{4}$$

3.3 Segmentation Loss

As mentioned above, our network contains two segmentation networks to solve the unsupervised domain adaptation problem. As shown in Fig. 1, in the forward process, E_s + S_s is our first segmentation network by training the source domain image x_s, E_t + S_t is the second segmentation network obtained by training the generated image $x_{s\to t}$. The loss function of two segmentation network is:

$$L_{seg}\left(E_s, S_s, E_t, S_t\right) = E_{x_s\sim X_s}\left[-y_s \log\left(S_s\left(E_s\left(x_s\right)\right)\right)\right] + E_{x_{s\to t}\sim X_{s\to t}}\left[-y_s \log\left(S_t\left(E_t\left(x_{s\to t}\right)\right)\right)\right] \tag{5}$$

As for backward process, segmentation network is employed in the test procedure to get the final segmentation result of the target domain image x_t. First, feeding target domain data x_t into the joint network G_2 through the segmentation branch E_t + S_t in the joint network for getting the first segmentation prediction $\hat{y}_t$. At the same time, $x_{t\to s}$ is generated by the translation branch E_t+$T_{t\to s}$. Then $x_{t\to s}$ will be fed into the joint network G_1, so the segmentation branch E_s+S_s of G_1 can output the second segmentation prediction for x_t, denoted as $\hat{y}_{t\to s}$. Finally, we fuse $\hat{y}_t$ and $\hat{y}_{t\to s}$ to get the final segmentation result $fusion\left\{\hat{y}_t, \hat{y}_{t\to s}\right\}$.

3.4 Feature-adaptation Loss

In order to further improve the segmentation results, we introduce feature adaptation loss referring to [Chen *et al.*, 2019]. As mentioned above, the segmentation network E_s + S_s is trained on x_s and used to induce $x_{t\to s}$. But there is still domain gap between them that makes the segmentation result of

Method	Dice Score [%]					Hausdorff Distance [mm]				
	AA	LAC	LVC	MYO	Average	AA	LAC	LVC	MYO	Average
without domain adaptation	28.4	27.7	4.0	8.7	17.2	32.6	35.1	54.2	57.4	44.8
CycleGAN[Zhu *et al.*, 2017]	73.8	75.7	52.3	28.7	57.6	16.2	15.4	20.8	27.7	20.0
CyCADA [Hoffman *et al.*, 2017]	72.9	**77.0**	62.4	45.3	64.4	14.9	12.8	18.3	22.9	17.2
AdaptSegNet[Tsai *et al.*, 2018]	69.4	73.3	54.9	36.7	58.6	17.1	17.6	22.8	25.4	20.7
SIFA[Chen *et al.*, 2019]	81.1	76.4	75.7	58.7	73.0	8.2	**10.5**	12.2	17.9	12.2
DSFN(our proposed)	**84.7**	76.9	**79.1**	**62.4**	**75.8**	**7.4**	11.9	**10.6**	**15.7**	**11.4**

Table 2: Comparison of Dice score and Hausdorff distance between our proposed method and other state-of-the-art methods for cardiac structures segmentation on MM-WHS dataset.

$x_{t\to s}$ inaccurate. To reduce their differences in feature level, we introduce a discriminator D_f^s and an auxiliary task for D_t, when D_f^s cannot distinguish between segmentation predictions from x_s or $x_{t\to s}$, and D_t can't differentiate between $x_{s\to t}$ and $x_{t\to s\to t}$, the features of x_s and $x_{t\to s}$ are consistent. The corresponding loss function is:

$$\begin{aligned} L_{adv}^{f_s}\left(E_s,S_s,D_f^s\right) =& E_{x_s\sim X_s}\left[logD_f^s\left(S_s\left(E_s\left(x_s\right)\right)\right)\right] \\ &+E_{x_{t\to s}\sim X_{t\to s}}\left[log\left(1-D_f^s\left(S_s\left(E_s\left(x_{t\to s}\right)\right)\right)\right)\right] \end{aligned} \tag{6}$$

and

$$\begin{aligned} L_{adv}^{t_{aux}}\left(E_s,D_t\right) =& E_{x_s\sim X_s}\left[logD_t\left(T_{s\to t}\left(E_s\left(x_s\right)\right)\right)\right] \\ &+E_{x_{t\to s}\sim X_{t\to s}}\left[log\left(1-D_t\left(T_{s\to t}\left(E_s\left(x_{t\to s}\right)\right)\right)\right)\right] \end{aligned} \tag{7}$$

Similarly, to reduce the domain gap between x_t and $x_{s\to t}$, we also introduce a discriminator D_f^t and an auxiliary task for D_s. And the loss function is:

$$\begin{aligned} L_{adv}^{f_t}\left(E_t,S_t,D_f^t\right) =& E_{x_{s\to t}\sim X_{s\to t}}\left[logD_f^t\left(S_t\left(E_t\left(x_{s\to t}\right)\right)\right)\right] \\ &+E_{x_t\sim X_t}\left[log\left(1-D_f^t\left(S_t\left(E_t\left(x_t\right)\right)\right)\right)\right] \end{aligned} \tag{8}$$

and

$$\begin{aligned} L_{adv}^{s_{aux}}\left(E_t,D_s\right) =& E_{x_{s\to t}\sim X_{s\to t}}\left[logD_s\left(T_{t\to s}\left(E_t\left(x_{s\to t}\right)\right)\right)\right] \\ &+E_{x_t\sim X_t}\left[log\left(1-D_s\left(T_{t\to s}\left(E_t\left(x_t\right)\right)\right)\right)\right] \end{aligned} \tag{9}$$

In summary, the overall objective function is:

$$\begin{aligned} L =& L_{adv}(E_s,T_{s\to t},D_t)+L_{adv}(E_t,T_{t\to s},D_s) \\ &+\lambda_{rec}L_{rec}\left(E_s,T_{s\to t},E_t,T_{t\to s}\right) \\ &+\lambda_{rec}L_{rec}\left(E_s,T_{s\to t},E_t,T_{t\to s}\right) \\ &+\lambda_f L_{adv}^{f_s}\left(E_s,S_s,D_f^s\right)+\lambda_f L_{adv}^{f_t}\left(E_t,S_t,D_f^t\right) \\ &+\lambda_{aux}L_{adv}^{t_{aux}}\left(E_s,D_t\right)+\lambda_{aux}L_{adv}^{s_{aux}}\left(E_t,D_s\right) \\ &+\lambda_{seg}L_{seg}\left(E_s,S_s,E_t,S_t\right) \end{aligned} \tag{10}$$

where we set $\lambda_{rec}=10$, $\lambda_f=0.1$, $\lambda_{aux}=0.1$ and $\lambda_{seg}=1$ to balance different losses.

3.5 Network Structure and Training Details

Our translation-segmentation joint network follows the structure settings from [Chen *et al.*, 2019]. Overall, our joint network contains an encoder connecting to two decoder branches. The encoder contains three residual blocks[He *et al.*, 2016] and down-sampling layers in sequence, then 8 residual blocks are followed and two dilated residual blocks continue to enlarge the receptive field, and two convolutional layers are connected at the end. For the translation branch, decoder contains a convolutional layer, 4 residual blocks and 3 deconvolutional layers to restore the resolution, then a convolutional layer and tanh activation function is followed. As for the segmentation branch, it contains a convolutional layer and an upsampling layer.

For the discriminator, we follow the setting of patchGAN[Isola *et al.*, 2017] and use five convolutional layers to convolve in turn, with the first three stride being 2 and the last two being 1. Their kernel size is 4×4 and the channel number is 64, 128, 256, 512 and 1, respectively.

During the training process, we substitute the least-squares loss for the log likelihood objective in the adversarial loss referring to CycleGAN[Zhu *et al.*, 2017], which can make the training process more stable[Mao *et al.*, 2017]. We use the Adam solver[Kingma and Ba, 2014] with a batch size of 8 and initialize the learning rate to 0.0002 for translation task and 0.001 for segmentation task. The various parts of the network are trained in the following order: E_s+S_s → $T_{s\to t}$ → D_t → D_f^s → E_t+S_t → $T_{t\to s}$ → D_s → D_f^t .

4 Experiments

We validate our proposed method on two datasets – Multi-Modality Brain Tumor Segmentation Challenge 2018[Menze *et al.*, 2014] and the Multi-Modality Whole Heart Segmentation Challenge 2017[Zhuang and Shen, 2016]. The first dataset contains four modalities of MRI imaging: T1, T1CE, T2, and FLAIR. There are a total of 75 patient data. Because experts always annotate whole tumor on T2 modality, we take T2 as the source domain with whole tumor annotation and aims to segment three target domain – T1, T1CE, FLAIR. Our experiments are conducted in an unpaired manner.

The second dataset is composed of 20 unpaired MR and CT imaging data, and their ground truth contains the annotation of four heart structures – the ascending aorta (AA), the left atrium blood cavity (LAC), the left ventricle blood cavity (LVC), and the myocardium of the left ventricle (MYO). We use MR images and their corresponding labels as the source domain dataset to segment the target domain CT images.

For both datasets, we randomly select 80% patient data for each modality as the training set and 20% as the testing set. All data subtract their mean and are divided by their standard deviation for normalization, then we switch them to [-1,1] before feeding to the network. We augment the data by rotation, cropping, etc., and each slice is resized to 256×256. The target labels of these two datasets are only used for evaluation.

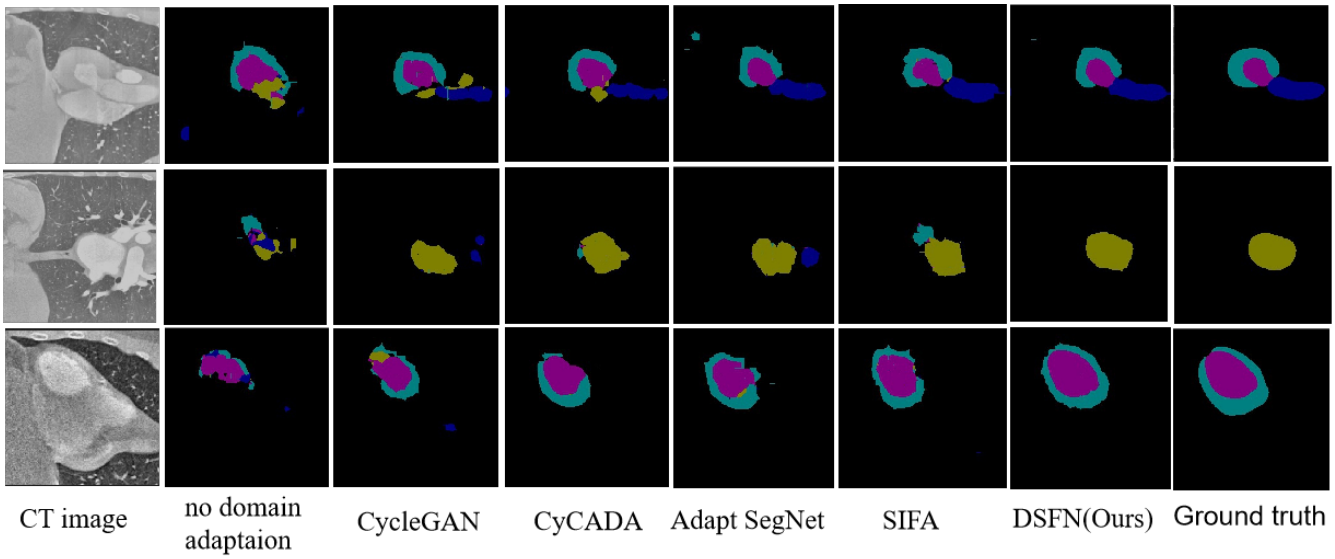


Figure 3: Visual display of the cross-domain cardiac structures segmentation results for comparing different methods on MM-WHS dataset.

Method	Dice Score [%]		
	Result1	Result2	Fusion
Without feature adaptation	60.6	61.8	64.2
With feature adaptation	62.5	63.4	**66.1**

Table 3: Ablation study of fusion strategy and feature adaptation on BRATS dataset. Result1 and Result2 represent the results of these two schemes, respectively.

Method	Dice Score [%]		
	Result1	Result2	Fusion
Without feature adaptation	58.3	60.1	64.7
With feature adaptation	70.9	73.7	**75.8**

Table 4: Ablation study of fusion strategy and feature adaptation on MM-WHS dataset. Result1 and Result2 represent the results of these two schemes, respectively.

Our experiments are conducted from two aspects: comparison with other state-of-the-art unsupervised domain adaptation methods and ablation experiments. For evaluation, we select two metrics commonly used in medical image segmentation tasks: Dice coefficient[%] and Hausdorff Distance. Dice coefficient measures the overlapping part of our prediction results and the groundtruth. Hausdorff distance is a distance defined between two sets in the metric space.

4.1 Comparison Experiments

The methods we compare with are: CycleGAN[Zhu *et al.*, 2017], CyCADA [Hoffman *et al.*, 2017], AdaptSegNet[Tsai *et al.*, 2018], SIFA[Chen *et al.*, 2019]. In addition, we also compare the results without domain adaptation (directly input the target image into the trained source domain segmentation network).

Table 1 and Table 2 respectively show the comparison result between our method and others on the brain tumor segmentation dataset and cardiac segmentation dataset. It can be seen that our method has a significant improvement compared to no domain adaptation result by improving Dice Score from 22.6% to 66.1% on tumor segmentation task and from 17.2% to 75.8% on cardiac structures segmentation task, and reducing the Hausdorff Distance from 46.8mm to 15.6mm and from 44.8mm to 11.4mm respectively, which shows that our method is very effective in solving domain degradation problem. More importantly, our proposed method achieves better results than the current state-of-the-art methods on these two datasets. Especially when compared with SIFA[Chen *et al.*, 2019], which only consider the source-to-target direction, we can increase the Dice by 6.8% on the BRATS dataset and by 2.8% on the MM-WHS datasetet after adopting both source-to-target and target-to-source direction and fusing their results, demonstrating the effectiveness of our proposed method.

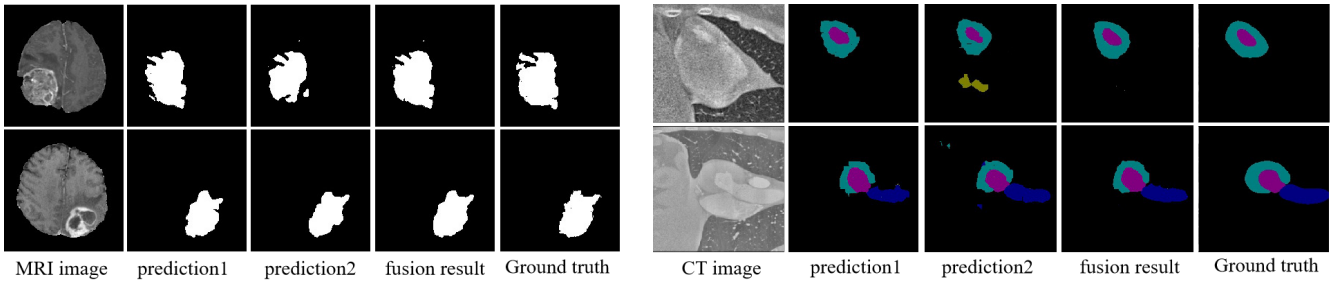
Fig.2 and Fig.3 present the visual results compared with other methods. We can note that our outputs are not only closer to the ground truth, but also the wrong semantic prediction results are far less than others, indicating that our fusion strategy can indeed make the results more robust.

4.2 Ablation Experiments

We perform ablation experiments on feature adaptation and fusion to verify the influence of each component for our proposed Dual-Scheme Fusion Network.

The ablation experiments are set as follows, we conduct experiments respectively with or without feature adaptation, in each case, we compare the results of three schemes: source-to-target scheme, target-to-source scheme, and their fusion solution. Table 3 and Table 4 present the ablation study results, Result1 represents the target-to-source method, Result2 represents the source-to-target method and Fusion represents our Dual-Scheme Fusion method.

It can be seen that, in both datasets, the results of adding



(a) Ablation study on BRATS

(b) Ablation study on MM-WHS

Figure 4: Ablation studies on (a) BRATS dataset and (b) MM-WHS dataset. The first and second rows present the results without and with feature adaptation, respectively. The second and third columns are the segmentation prediction results of two schemes, and the fourth column is their fusion result, respectively.

feature adaptation are improved in all three cases compared with no feature adaptation. Among them, the improvement on the MM-WHS dataset is particularly obvious, and the Dice score can increase over 10% compared to no feature adaptation, demonstrating the effectiveness of feature adaptation.

And regardless of whether feature adaptation is added or not, compared with experiments conducted from source-to-target or target-to-source perspective alone, the experimental results are more accurate after using our dual-scheme fusion strategy on both datasets. And Dice score can improve about 3% on BRATS dataset and 4% on MM-WHS dataset without feature adaptation. If feature adaptation is added, our method can further improve the Dice score by about 3% on BRATS and 2% on MM-WHS, which reflects the effectiveness of fusion strategy.

Fig.4(a) and Fig.4(b) present the visual results on BRATS dataset and MM-WHS dataset respectively. The first row shows the results without feature adaptation while the second row shows them with feature adaptation. It can be seen that overall the results of adding feature adaptation are more accurate than not adding. The first column is the input images, prediction1 is the results from target-to-source scheme, prediction2 is from source-to-target scheme and fusion result means adopting our fusion strategy. We can see that the edges of fusion images are smoother than the two separate prediction results and are closer to the ground truth. Moreover, from Fig.4(b) we can see that the misprediction outside the cardiac structures from the two prediction results also disappear after fusion strategy, which means through our fusion strategy, the results from target-to-source and source-to-target perspectives can complement each other and a smoother and more robust result can be obtained.

5 Conclusion

In this paper, we address the problem of domain adaptation from two complementary perspectives and unify these two approaches into an end-to-end network in an elegant and efficient way. Through the cycle structure, not only can the unpaired data training problem be solved, but also a clever combination of these two perspectives is obtained. Finally, we choose an effective fusion method to acquire the final result. For further improvement, we also add feature adaptation loss. The experimental results show that our proposed method greatly surpasses the baseline and is superior to other state-of-the-art methods.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants 62041105, 6182211, the Science and Technology Major Project of Hubei Province (Next-Generation AI Technologies) under Grant 2019AEA170, Natural Science Foundation of Hubei Province under Grants 2018CFA050. The numerical calculations in this paper have been done on the supercomputing system in the Supercomputing Center of Wuhan University.

References

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

[Chen *et al.*, 2018] Cheng Chen, Qi Dou, Hao Chen, and Pheng-Ann Heng. Semantic-aware generative adversarial nets for unsupervised domain adaptation in chest x-ray segmentation. In *International Workshop on Machine Learning in Medical Imaging*, pages 143–151. Springer, 2018.

[Chen *et al.*, 2019] Cheng Chen, Qi Dou, Hao Chen, Jing Qin, and Pheng-Ann Heng. Synergistic image and feature adaptation: Towards cross-modality domain adaptation for medical image segmentation. *arXiv preprint arXiv:1901.08211*, 2019.

[Dou *et al.*, 2018] Qi Dou, Cheng Ouyang, Cheng Chen, Hao Chen, and Pheng-Ann Heng. Unsupervised cross-modality domain adaptation of convnets for biomedical image segmentations with adversarial loss. *arXiv preprint arXiv:1804.10916*, 2018.

[Fang and Yan, 2020] Xi Fang and Pingkun Yan. Multi-organ segmentation over partially labeled datasets

with multi-scale feature abstraction. *arXiv preprint arXiv:2001.00208*, 2020.

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

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[Hoffman *et al.*, 2017] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. *arXiv preprint arXiv:1711.03213*, 2017.

[Isola *et al.*, 2017] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

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

[Kingma and Ba, 2014] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[Long *et al.*, 2015] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I Jordan. Learning transferable features with deep adaptation networks. *arXiv preprint arXiv:1502.02791*, 2015.

[Mao *et al.*, 2017] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *ICCV*, pages 2794–2802, 2017.

[Menze *et al.*, 2014] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014.

[Nie *et al.*, 2017] Dong Nie, Roger Trullo, Jun Lian, Caroline Petitjean, Su Ruan, Qian Wang, and Dinggang Shen. Medical image synthesis with context-aware generative adversarial networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 417–425. Springer, 2017.

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

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

[Wolterink *et al.*, 2017] Jelmer M Wolterink, Anna M Dinkla, Mark HF Savenije, Peter R Seevinck, Cornelis AT van den Berg, and Ivana Išgum. Deep mr to ct synthesis using unpaired data. In *International Workshop on Simulation and Synthesis in Medical Imaging*, pages 14–23. Springer, 2017.

[Zhao *et al.*, 2018] He Zhao, Huiqi Li, Sebastian Maurer-Stroh, Yuhong Guo, Qiuju Deng, and Li Cheng. Supervised segmentation of un-annotated retinal fundus images by synthesis. *IEEE transactions on medical imaging*, 38(1):46–56, 2018.

[Zhu *et al.*, 2017] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, pages 2223–2232, 2017.

[Zhu *et al.*, 2020] Qikui Zhu, Bo Du, and Pingkun Yan. Boundary-Weighted Domain Adaptive Neural Network for Prostate MR Image Segmentation. *IEEE Transactions on Medical Imaging*, 39(3):753–763, March 2020.

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