Enhancing Unsupervised Domain Adaptation via Semantic Similarity Constraint for Medical Image Segmentation

Tao Hu , **Shiliang Sun**[*] , **Jing Zhao** and **Dongyu Shi**
School of Computer Science and Technology, East China Normal University, Shanghai 200062, China
slsun@cs.ecnu.edu.cn

Abstract

This work proposes a novel unsupervised cross-modality adaptive segmentation method for medical images to tackle the performance degradation caused by the severe domain shift when neural networks are being deployed to unseen modalities. The proposed method is an end-2-end framework, which conducts appearance transformation via a domain-shared shallow content encoder and two domain-specific decoders. The feature extracted from the encoder is directly enhanced to be more domain-invariant by a similarity learning task using the proposed Semantic Similarity Mining (SSM) module which has a strong help of domain adaptation. The domain-invariant latent feature is then fused into the target domain segmentation sub-network, trained using the original target domain images and the translated target images from the source domain in the framework of adversarial training. The adversarial training is effective to narrow the remaining gap between domains in semantic space after appearance alignment. Experimental results on two challenging datasets demonstrate that our method outperforms the state-of-the-art approaches.

1 Introduction

Medical image segmentation is an important research area, and deep learning is successful in many related tasks with this area. Due to different physical imaging principles and imaging modes, the appearance of achieved digital images will have significant differences, and domain shift is even more severe than natural images. As shown in Figure 1, the magnetic resonance imaging (MRI) and computed tomography (CT) images of the same cardiac region produced markedly different appearances. Well-trained deep neural networks on one domain often fail when being deployed on another domain of the same target. It is necessary to transfer the knowledge learned from one domain to another domain for reducing the performance degradation caused by domain shift. Many efforts have been taken to reduce the performance degradation caused by domain shifting. The unsupervised domain adaptation approach is a solution for it.

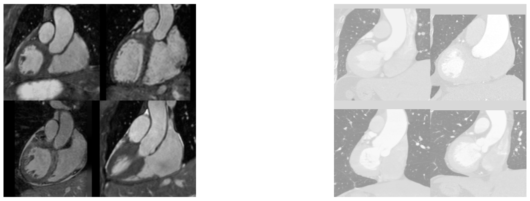
(a) Examples of MR images (b) Examples of CT images

Figure 1: Image examples from CT and MR domains which can illustrate the severe cross-modality domain shift of medical images.

Previous works tackle this problem from three main aspects. One is the image-level adaptation, which reduces domain shift by transforming the images between domains in appearance at pixel-level. These methods commonly adopt the cycle-consistency constraint proposed in CycleGAN [Zhu *et al.*, 2017], which can translate images between unpaired images from different distributions. The second aspect is adapting latent features. The goal of these approaches is to extract domain-invariant features through deep neural networks. Most of these methods are implemented by deploying adversarial training on latent feature distributions. [Tsai *et al.*, 2018] proposed to add discriminator into a more compact space, such as the segmentation mask space. The third aspect is combining the above two perspectives to adapt in image level and feature level simultaneously.

The methods mentioned above perform adaptation via cycle-consistency loss and adversarial learning, which have some disadvantages: 1) if a discriminator is added to a high-dimensional and over-complex feature space, it could lead to non-convergence in the GAN training process; 2) domain-invariant semantic feature information could be destructed because of the instability of adversarial training in the segmentation mask space. To solve these problems, we design a novel network structure that can take full advantage of the adversarial learning without destructing the semantic information, and we directly enhance the extracted domain-invariant features via our proposed additional semantic similarity mining module. Our method supports synergistic adaptations from both image and feature spaces simultaneously. The pro-

[*] Corresponding Author.

posed method has two workflows. Specifically, in the first flow, we use a shallow content encoder shared between domains and two domain-specific decoders to translate images between the two domains towards appearance but keep the semantic content. The design of the encoder is shallow because the weight of the encoder is shared by the two modalities. If the content encoder has too much capacity, it would not capture the domain invariant feature but be equivalent to using two independent domain encoders. The image translation workflow is constrained by cycle consistency and adversarial learning. Furthermore, to enhance the extracted domain-invariant feature after the content encoder, we add a strong semantic similarity mining (SSM) task on the content encoder. This task gives the content encoder a strict constraint that the features extracted by the encoder of two augmented images from the same original image should be similar. Another flow uses the synthesis target domain images from the source domain and the original target domain images to train a segmentation model. An adversarial loss is added to the segmentation prediction space between real images and synthesis images. Our experiments show that the proposed method outperforms the state-of-the-art (SOTA) methods with a large margin. The major contributions of this work are listed as below:

- We propose a novel network structure to address the unsupervised domain adaptation problem for medical image segmentation.
- This is the first attempt to add semantic similarity mining to the UDA learning framework to directly enhance the extracted domain-invariant features. The semantic similarity mining task is model-independent and could be easily integrated to other models.
- We evaluate the proposed method on two challenging multi-modal medical structure segmentation datasets, and the proposed method outperforms previous state-of-the-art methods by a large margin.

2 Related Work

Unsupervised domain adaptation is an active research direction recently because it does not require the annotation data of the target domain. It has received significant attention in the field of medical image processing cause heterogeneous domain shift is more severe due to different physical principles. Previous work can be roughly divided into three streams: only feature, only image, image + feature.

1) Feature Alignment. The aim of this stream is to learn domain-invariant semantic features between domains. Most of the models follow the Domain Adversarial Neural Network (DANN) and Adversarial Discriminative Domain Adaptation (ADDA) structure [Ganin *et al.*, 2016; Tzeng *et al.*, 2017]. [Dou *et al.*, 2019] assumed that the severe data shift across domains mainly exists in low-level image features. They proposed to fix the weights of high-level layers in the network and adapt the low-level layers between domains via adversarial learning, and they evaluated the method on the MMWHS dataset [Zhuang and Shen, 2016]. A disadvantage of this method is that it needs to manually select the hyperparameter n, which defines the number of low-level layers.

2) Image Alignment. Driven by the mighty generative power of Generative Adversarial Network [Goodfellow *et al.*, 2014], a lot of works have been proposed in image alignment. [Bousmalis *et al.*, 2017] proposed to translate source domain images to the target domain, then use the target-like images for supervised training, which have inherited labels from the source domain. Inversely, [Zhang *et al.*, 2018b] use the labeled source domain data for supervised training and then transform target domain data to source domain for testing using the pre-trained source model.

3) Image + Feature Alignment. As previously mentioned, the two alignment perspectives are fundamental, and they are complementary to each other in fact. Some recent studies combined these two perspectives to obtain a stronger model for unsupervised domain adaptation [Chen *et al.*, 2020a]. In [Hoffman *et al.*, 2018] and [Zhang *et al.*, 2018a], their methods adapted image-level and feature-level space between synthetic and real-world driving scene domains separately and connected the two stages sequentially, so the internal interactions between the two perspectives can not be achieved. [Chen *et al.*, 2020b] proposed a framework named Synergistic Image and Feature Alignment (SIFA) to conduct synergistic image and feature alignment for cross-modality cardiac image segmentation.

3 Approach

Figure 2 shows the overview of the proposed method. The framework consists of several neural network modules described as follow, a content encoder E_c, two upsampling style decoders $\{U_s, U_t\}$, a segmentation encoder E_s, a pixel classifier C for the final segmentation output, and two MLP blocks.

3.1 Image and Feature Space Alignment

Denote $\{x^s, y^s\}$ are samples from the labeled source domain $\mathcal{X}^s$ and $\{x^t\}$ are samples from the unlabeled target domain $\mathcal{X}^t$. There are two appearance translation sub-networks which are $E_c \circ U_s$ and $E_c \circ U_t$ who can translate images cross domains or reconstruct images towards appearance. The content encoder E_c is designed not very deep because its weight is shared between two modalities. If the encoder has too much capacity, it would not capture the domain invariant feature but would be equivalent to using two independent domain encoders. The workflows of image translation is shown in Figure 2.

A reconstruction loss is defined to ensure the generated images can preserve the content and style information. The reconstruction loss is defined as:

$$\mathcal{L}_{rec} = \frac{1}{2}||x^{s\to s} - x^s||_1 + \frac{1}{2}||x^{t\to t} - x^t||_1 \tag{1}$$

To ensure the synthesis images $x^{s\to t}$ and $x^{t\to s}$ to preserve the original semantic content information, we adopt to use the commonly-used cycle-consistency loss [Zhu *et al.*, 2017] for unpaired cross-modality images alignment:

$$\mathcal{L}_{cyc} = \frac{1}{2}||x^{s\to t\to s} - x^s||_1 + \frac{1}{2}||x^{t\to s\to t} - x^t||_1 \tag{2}$$

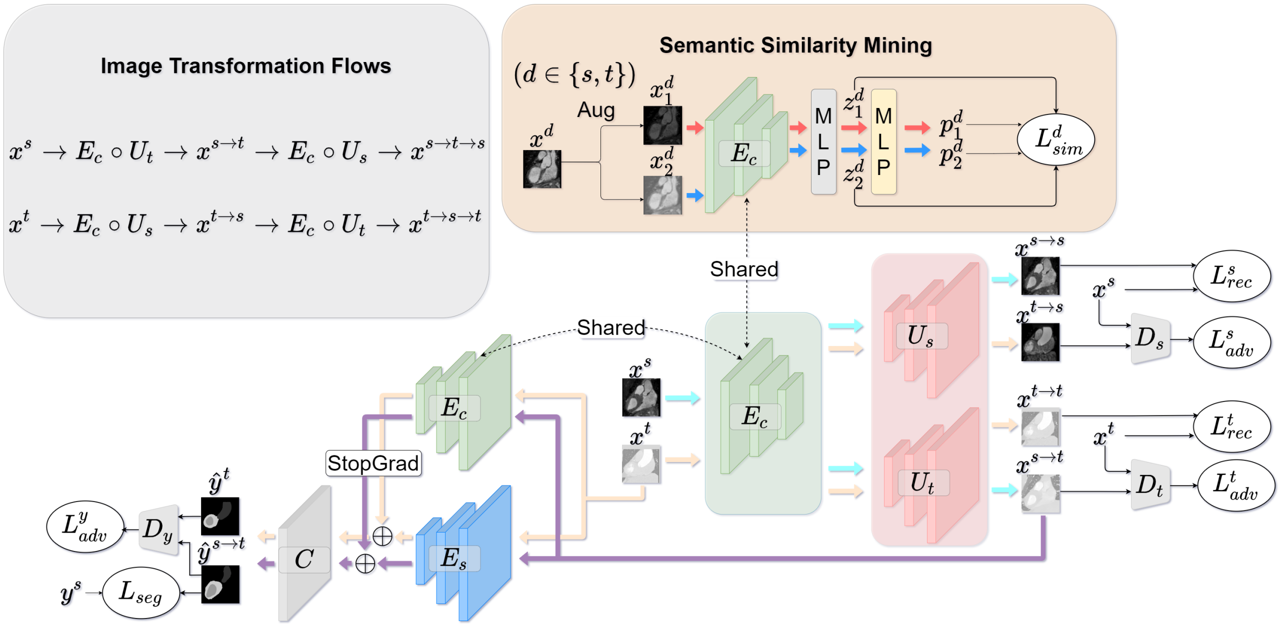

Figure 2: Overview of the proposed method. The E_c, U_s, and U_t constitute the image appearance transformation module which performs image transformation towards appearance between the source and the target domain. The segmentation encoder E_s is trained using real target domain images and synthesis target-like images. An additional semantic similarity mining task is added on the content encoder E_c to enhance the semantics extraction. The discriminators $\{D_s, D_t, D_y\}$ differentiate their inputs to derive adversarial losses.

According to CycleGAN [Zhu *et al.*, 2017], the adversarial losses are added on the synthesis images $x^{s\rightarrow t}$ and $x^{t\rightarrow s}$ to ensure them like the realistic images:

$$\begin{aligned}\mathcal{L}_{adv} =&\mathbb{E}_{x^s\sim\mathcal{X}^s}[logD_s(x^s)]+\\&\mathbb{E}_{x^t\sim\mathcal{X}^t}[log(1-D_s(G_s(x^t)))]+\\&\mathbb{E}_{x^t\sim X^t}[logD_t(x^t)]+\\&\mathbb{E}_{x^s\sim X^s}[log(1-D_t(G_t(x^s)))]\end{aligned} \tag{3}$$

In above equations, G_s equals to $E_c \circ U_t$, G_t equals to $E_c \circ U_s$. $D_{\{s,t\}}$ means the corresponding discriminator.

With the constraints of cycle-consistency loss and adversarial loss, the latent feature $f = E_c(x^{\{s,t\}})$ will preserve some semantic content information. At the same time, the distribution of synthesis target domain image $x^{s\rightarrow t}$ is bringing closer to the real data distribution of the target domain, which can be used to train a target domain segmentation network in a supervised learning manner.

3.2 Segmentation Sub-network

We use a target segmentation encoder E_s which accepts input of target-like images to narrow the remaining gap between target domain images x^t and synthesis target domain images $x^{s\rightarrow t}$. Caused by the instability of GAN training, the generated target-like image $x^{s\rightarrow t}$ could lose some semantic information. To tackle this problem, we integrate the content encoder E_c into the segmentation sub-module. The $x^{s\rightarrow t}$ and x^t are forwarded into the segmentation encoder E_s directly and then the feature extracted by the segmentation encoder E_s will be fused with the feature extracted by the content encoder E_c. Finally, a pixel-wise segmentation classifier is applied to predict the final segmentation mask using the information from the content encoder E_c and the segmentation encoder E_s. In this part, supervised learning is applied to learn the prediction of segmentation mask using synthesis target-like images $x^{s\rightarrow t}$ with label y^s inherited from x^s. The segmentation mask predicted from $x^{\{s,t\}}$ can be denoted by

$$\hat{y}^{\{s,t\}} = C(E_c(x^{\{s,t\}}) \oplus E_s(x^{\{s,t\}})) \tag{4}$$

We implement the segmentation task learning by minimizing a hybrid loss L_{seg}:

$$\mathcal{L}_{seg} = Dice(\hat{y}^{s\rightarrow t}, y^s) + FL(\hat{y}^{s\rightarrow t}, y^s) + T(\hat{y}^{s\rightarrow t}, y^s) \tag{5}$$

In above equation, Dice means the widely used Dice loss. We implement Dice loss with Laplacian smoothing, which is defined as below:

$$Dice(X,Y) = 1 - \frac{2|X\cap Y|+1}{|X|+|Y|+1} \tag{6}$$

The FL term means Focal loss [Lin *et al.*, 2017]. It can be regarded as a kind of extension of cross-entropy loss and it is mainly used to tackle the unbalanced samples problem.

$$FL(p,y) = \begin{cases}-\alpha(1-p)^{\gamma}log(p) & \text{if } y=1,\\ -(1-\alpha)p^{\gamma}log(1-p) & \text{if } y=0.\end{cases} \tag{7}$$

where α and γ are hyper-parameters. In our experiments, corresponding values are setting to 0.75 and 2, respectively.

The term T in Equation (5) is the Tversky loss [Salehi *et al.*, 2017]:

$$T(A,B) = 1 - \frac{A\cap B}{A\cap B + \alpha|A-B| + \beta|B-A|} \tag{8}$$

where α and β are hyper-parameters, both set to 0.5.

Similarly, adversarial learning is applied to $\hat{y}^t$ and $\hat{y}^{s\rightarrow t}$.

$$\begin{aligned}\mathcal{L}^y_{adv}(S, D_y) =& \mathbb{E}_{\hat{y}^t}[logD_y(\hat{y}^t)]+ \\ & \mathbb{E}_{\hat{y}^{s\rightarrow t}}[log(1 - D_y(S(\hat{y}^{s\rightarrow t})))]\end{aligned} \tag{9}$$

where S equals to $C \circ (E_c \oplus E_s)$.

3.3 Additional Constraint via Semantic Similarity Mining

The domain invariant representation is learned from image reconstruction loss, cycle-consistency loss, and adversarial loss. However, these constraints only indirectly affect the domain invariant representation learning. We introduce adding a strong semantic similarity mining (SSM) sub-task into our framework to directly enhance the domain invariance of the feature extracted by the content encoder E_c.

This technique is based on the structure of Siamese networks [Chen and He, 2021]. Siamese networks accept two or more inputs to be processing by weight-sharing networks. The backbone of the Siamese networks is the content encoder E_c. A projection MLP head and a prediction MLP head (denoted as h) are connected to the end of the backbone. We set the encode function f consisting of the backbone and the projection MLP head and we denote $z = f(x)$ and $p = h(f(x))$ as the outputs of the encode function and the prediction MLP head given the input x. Given $\{z^k_i, p^k_i | i = 1, 2 \text{ and } k = s, t\}$ for the i-th random semantic preserved augmented image x^k (k represents domain), we minimize the cosine similarity loss L_{sim} by the below term:

$$\mathcal{D}(a, b) = 1 - \frac{a}{||a||_2}\frac{b}{||b||_2} \tag{10}$$

$$\begin{aligned}\mathcal{L}_{sim} = \frac{1}{2}\sum_{k=s,t} (&\mathcal{D}(p^k_1, stopgrad(z^k_2))+ \\ & \mathcal{D}(p^k_2, stopgrad(z^k_1)))\end{aligned} \tag{11}$$

where $||\cdot||_2$ is ℓ_2-norm, $stopgrad$ is the stop-gradient operation. The model maximizes the cosine-similarity of the outputs between two augmented images from one image.

The existing methods mainly use cycle-consistency and adversarial training to constrain domain-invariant feature learning, both of which are applied at the image level to **indirectly** guide the feature level alignment. We introduce our proposed semantic similarity mining (SSM) module to **directly** guide the learning of domain invariance from the deep feature level, which does not require as much data as training a GAN needed. We require the Siamese network to output as similar as possible between two different image-augmented versions of an original image from any modality. The augmentation method of images is well designed for specific modality types to confuse the image appearance between the modalities as much as possible, so we can use this method to mine the semantic similarity cross modalities which helps enhancing domain invariance a lot. Above mentioned random semantic preserved augmentation plays an essential role in the semantic similarity mining task. Every augmentation type in the augmentation set should be semantic preserved, which means the label would not change at all. The augmentation should apply randomly every time.

3.4 Overall Objective

The overall objective function is defined as follows:

$$\begin{aligned}\mathcal{L} =& \lambda_{rec}\mathcal{L}_{rec} + \lambda_{cyc}\mathcal{L}_{cyc} + \lambda_{adv}\mathcal{L}_{adv}+ \\ & \lambda_{seg}\mathcal{L}_{seg} + \lambda^y_{adv}\mathcal{L}^y_{adv} + \lambda_{sim}\mathcal{L}_{sim}\end{aligned} \tag{12}$$

The $\{\lambda_{rec}, \lambda_{cyc}, \lambda_{adv}, \lambda_{seg}, \lambda^y_{adv}, \lambda_{sim}\}$ in Equation (12) are trade-off parameters adjusting the weight of each component. The corresponding values are set to $\{10.0, 1.0, 1.0, 1.0, 0.01, 1.0\}$ respectively. The $\mathcal{L}_{rec}$, $\mathcal{L}_{cyc}$, $\mathcal{L}_{adv}$, and $\mathcal{L}_{seg}$ are similar to previous works. Qualitatively, it can be considered that the following three types of constraints are successively weakened: the reconstruction constraints $\mathcal{L}_{rec}$ > the domain alignment constraints $\mathcal{L}_{cyc}$ and $\mathcal{L}_{adv}$ > the segmentation realistic constraints $\mathcal{L}^y_{adv}$. So these hyper-parameters are set to 10,1,0.01 respectively. The last term $\mathcal{L}_{sim}$ is specific to our method, it is a loss term that also constrains the alignment between modalities, so it is set to 1, too. We tried other values like 10, 0.1, and our method performs best in the value of 1.

3.5 Optimization Steps

The entire training process is End-2-End, we do not train every module separately. But in each training iteration, we first optimize the image translation module and semantic similarity mining module together, then optimize the discriminators of source and target domains. After that, we update the parameters of the segmentation module. More training details and network structures could be found in the supplementary material.

4 Experiments

4.1 Dataset

Cardiac substructure segmentation. Multi-Modality Whole Heart Segmentation Challenge 2017 dataset (MMWHS) [Zhuang and Shen, 2016] is a cardiac segmentation dataset including two modality images (MR and CT). Each modality contains 20 volumes collected from different sites, and there is no pair relationship between modalities. We choose four classes of the cardiac structures. They are the ascending aorta (AA), the left atrium blood cavity (LAC), the left ventricle blood cavity (LVC), and the myocardium of the left ventricle (MYO). **Abdominal multi-organ segmentation.** We choose the public CT data from [Landman *et al.*, 2017] (30 volumes) and T2-SPIR MRI training data from the ISBI 2019 CHAOS Challenge [Selver *et al.*, 2019] (20 volumes) to validate our method. Ground truth segmentation masks of four abdominal organs including liver (L), right kidney (R.K), left kidney (L.K), and spleen (SP) are provided in both modalities.

For each dataset, we randomly split them with 70% cases for training, 10% cases for validation and 20% cases for test. For training our model, online augmentation is applied to reduce over-fitting, which mainly includes 2D/3D random elastic deformation, random rotation, random scaling, and 3D affine transformation. The final training data is 2D coronal slices in the resolution of 256x256x1 sampled from the augmented volumes.

Cardiac MR to Cardiac CT										
Method	Dice↑					ASD↓				
	AA	LAC	LVC	MYO	Average	AA	LAC	LVC	MYO	Average
Supervised training	92.7	91.1	91.9	87.7	90.9	1.5	3.5	1.7	2.1	2.2
W/o adaptation	28.4	27.7	4.0	8.7	17.2	20.6	16.2	N/A	48.4	N/A
PnP-AdaNet [Dou *et al.*, 2019]	74.0	68.9	61.9	50.8	63.9	12.8	6.3	17.4	14.7	12.8
AdaOutput [Tsai *et al.*, 2018]	65.2	76.6	54.4	43.6	59.9	17.9	5.5	5.9	8.9	9.6
CycleGAN [Zhu *et al.*, 2017]	73.8	75.7	52.3	28.7	57.6	11.5	13.6	9.2	8.8	10.8
CyCADA [Hoffman *et al.*, 2018]	72.9	77.0	62.4	45.3	64.4	9.6	8.0	9.6	10.5	9.4
SIFA [Chen *et al.*, 2020b]	81.3	79.5	73.8	61.6	74.1	7.9	6.2	5.5	8.5	7.0
DSFN [Zou *et al.*, 2021]	**84.7**	76.9	79.1	62.4	75.8	N/A	N/A	N/A	N/A	N/A
DSAN [Han *et al.*, 2021]	79.9	84.8	82.8	66.5	78.5	7.7	6.7	3.8	5.6	5.9
Ours	82.0	**85.3**	**88.4**	**67.6**	**80.8**	**6.2**	**4.1**	**3.0**	**3.4**	**4.2**

Abdominal MR to Abdominal CT										
Method	Dice↑					ASD↓				
	L	R.K	L.K	SP	Average	L	R.K	L.K	SP	Average
Supervised training	92.8	86.4	87.4	88.2	88.7	1.0	1.8	0.9	1.2	1.2
W/o adaptation	73.1	47.3	57.3	55.1	58.2	2.9	5.6	7.7	7.4	5.9
SynSeg-Net [Huo *et al.*, 2018]	85.0	82.1	72.7	81.0	80.2	2.2	1.3	2.1	2.0	1.9
AdaOutput [Tsai *et al.*, 2018]	85.4	79.7	79.7	81.7	81.6	1.7	1.2	1.8	**1.6**	1.6
CycleGAN [Zhu *et al.*, 2017]	83.4	79.3	79.4	81.7	81.6	1.8	1.3	**1.2**	1.9	1.8
CyCADA [Hoffman *et al.*, 2018]	84.5	78.6	80.3	76.9	80.1	2.6	1.4	1.3	1.9	1.8
SIFA [Chen *et al.*, 2020b]	88.0	**83.3**	80.9	82.6	83.7	**1.2**	**1.0**	1.5	**1.6**	**1.3**
Ours	**88.5**	**83.3**	**82.0**	**83.1**	**84.2**	1.3	**1.0**	**1.2**	**1.6**	**1.3**

Table 1: Quantitative comparison with different unsupervised domain adaptation methods for cardiac/abdominal segmentation.

Methods	Dice↑					ASD↓				
	AA	LAC	LVC	MYO	Average	AA	LAC	LVC	MYO	Average
W/o adaptation	28.4	27.7	4	8.7	17.2	20.6	16.2	N/A	48.4	N/A
Only image adaptation	73.4	75.7	49.2	35.7	58.5	10.5	11.5	8.9	8.5	9.9
Unshared encoder	68.9	72.1	72.0	59.6	71.4	7.9	6.2	4.5	4.4	5.8
Without SSM	74.8	82.1	86.3	64.5	76.9	7.8	4.5	3.5	3.5	4.8
Ours	**82.0**	**85.3**	**88.4**	**67.6**	**80.8**	**6.2**	**4.1**	**3.0**	**3.4**	**4.2**

Table 2: Effectiveness of each key component in our proposed method. The results are from Cardiac MR to CT adaptation

4.2 Evaluation Metrics

We choose the Dice similarity coefficient (Dice) and the average surface distance (ASD) as our evaluation metrics. Dice similarity coefficient is a measurement that measures the volume overlap between the prediction and corresponding ground truth. Average surface distance evaluates the distance between the surface of the prediction and the label. A higher Dice or a lower ASD means better performance.

4.3 Comparison with Other Methods

In this section, we compared our method with several state-of-the-art methods in quantitative and qualitative. PnP-AdaNet [Dou *et al.*, 2019] proposed a plug-and-play style method to align the feature spaces of the domains; AdaOutput [Tsai *et al.*, 2018] conducts an adversarial task between multi-level prediction maps; CycleGAN [Zhu *et al.*, 2017] performs an image-to-image translation by a cyclist reconstruction manner; CyCADA [Hoffman *et al.*, 2018] aligns the feature from different domains at both image and feature level; SIFA [Chen *et al.*, 2020b] conducts synergistic alignment of the domain from both image and feature perspectives; DSFN [Zou *et al.*, 2021] introduces a dual-scheme network structure to reduce the domain gap; DSAN [Han *et al.*, 2021] proposed a deep symmetric architecture to eliminate the domain gap. In the above-mentioned methods, AdaOutput, CycleGAN, and CyCADA are proposed for natural image datasets, but PnP-AdaNet, SIFA, DSFN and DSAN are proposed for medical images.

Table 1 shows the comparison on the two datasets in the adaptation direction of MR to CT. Supervised training result for this task is took as the performance upper bound, and the result that the model trained on the source domain but directly applied to the target domain is obtained as the lower bound. Both of them are recorded in the ”Supervised training W/o adaptation”. In the adaptation direction of cardiac MR to cardiac CT, according to the ”Supervised training W/o adaptation” result, we can obtain that the apparent differences caused by domain shift are severe, which drop the performance of average 90.9% to 17.2% in Dice coefficient. With our proposed unsupervised domain adaptation method, we achieve the average Dice coefficient of 80.8% and the ASD value of 4.2 remarkably. In the abdominal dataset, result is

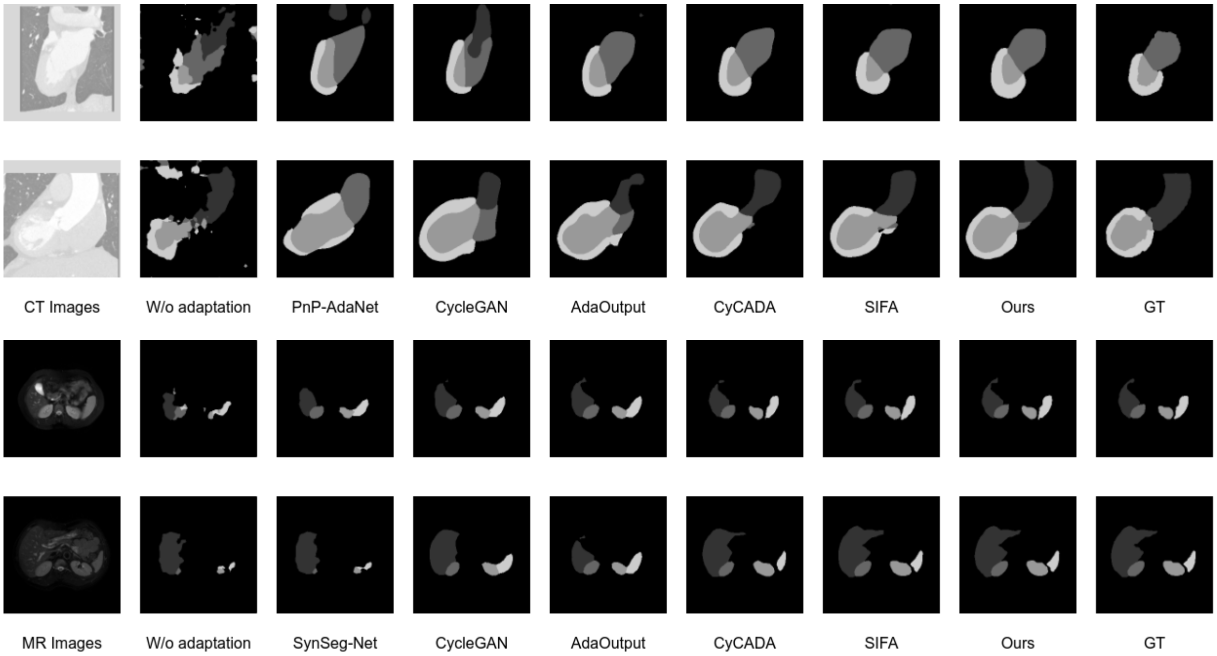

Figure 3: Visual comparison of the predicted segmentation masks of different methods.

similar but the task is more easy. Visual comparison results are provided in Figure 3. We can see that in the ”W/o adaptation” case, the prediction is definitely wrong. Our method outperforms all of the other methods. More experiment results are reported in supplementary material.

4.4 Ablation Study

We conducted a number of ablation studies to evaluate the effectiveness of each key component in our proposed model. The results are shown in Table 2.

The baseline method is using image adaptation only, this method train the image translation-related modules, and test on an extra segmentation network using target-like images. Compared to ”W/o adaptation” lower-bound, the result shows that this method effectively narrows the gap between domains because the Dice increased to 58.5% compared to the ”W/o adaptation” case.

Next, we study the necessity of sharing weights between encoders. We conduct a study that using two independent content encoders for every domain. In this case, the average Dice dropped 9.4%. It proves that sharing the weights of the content encoder between domains is necessary for extracting domain-invariant features.

We use a light-weight nine layers convolutional neural network as the content encoder, which is relatively shallower than the commonly used deep structures which usually have dozens of layers, and we also conduct experiments that change the number of layers in the encoder E_c. The average Dice comparison of different number of layers in the encoder E_c is shown in Table 3.

N of layers	5	9 (ours)	15	21	25
Avg. Dice	75.8	80.8	73.6	77.9	61.4

Table 3: Average Dice of different number of layers in the encoder.

The semantic similarity loss is an important part of our method, and we also study its effectiveness. In comparison, we remove this loss term in the overall loss term and train again. We can see a reduction of 3.9% in the average Dice compared to the original model.

5 Conclusion

We proposed a novel unsupervised domain adaptation medical image segmentation algorithm that combines both unpaired image transformation and domain invariant feature alignment learning. Domain invariant features are directly enhanced by the proposed SSM module. Extensive experiments with promising results on two challenging datasets show the effectiveness of our approach. At present, all of the experiments only use 2D volume slices, and it is also possible to extend our method to 3D volumes. From a practical perspective, our method has the potential to be extended to evaluate the quality of cross-domain image translation, e.g., via scoring the segmentation predictions of real and translated images by the discriminator. These could be explored in our future work.

Acknowledgments

This work was supported by the NSFC Projects 62076096 and 62006078, the Shanghai Municipal Project 20511100900, Shanghai Knowledge Service Platform Project ZF1213, STCSM Project 22ZR1421700 and the Fundamental Research Funds for the Central Universities.

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 and He, 2021] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021.

[Chen *et al.*, 2020a] Cheng Chen, Q. Dou, Hao Chen, J. Qin, and P. Heng. Unsupervised bidirectional cross-modality adaptation via deeply synergistic image and feature alignment for medical image segmentation. *IEEE Transactions on Medical Imaging*, 39:2494–2505, 2020.

[Chen *et al.*, 2020b] Cheng Chen, Qi Dou, Hao Chen, Jing Qin, and Pheng Ann Heng. Unsupervised bidirectional cross-modality adaptation via deeply synergistic image and feature alignment for medical image segmentation. *IEEE transactions on medical imaging*, 39(7):2494–2505, 2020.

[Dou *et al.*, 2019] Qi Dou, Cheng Ouyang, Cheng Chen, Hao Chen, Ben Glocker, Xiahai Zhuang, and Pheng-Ann Heng. Pnp-adanet: Plug-and-play adversarial domain adaptation network at unpaired cross-modality cardiac segmentation. *IEEE Access*, 7:99065–99076, 2019.

[Ganin *et al.*, 2016] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

[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. *Advances in neural information processing systems*, 27, 2014.

[Han *et al.*, 2021] Xiaoting Han, Lei Qi, Qian Yu, Ziqi Zhou, Yefeng Zheng, Yinghuan Shi, and Yang Gao. Deep symmetric adaptation network for cross-modality medical image segmentation. *arXiv preprint arXiv:2101.06853*, 2021.

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

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

[Landman *et al.*, 2017] Bennett Landman, Z Xu, J Igelsias, M Styner, T Langerak, and A Klein. Multi-atlas labeling beyond the cranial vault-workshop and challenge. *Accessed: Jul*, 2017.

[Lin *et al.*, 2017] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[Salehi *et al.*, 2017] Seyed Sadegh Mohseni Salehi, Deniz Erdogmus, and Ali Gholipour. Tversky loss function for image segmentation using 3d fully convolutional deep networks. In *International workshop on machine learning in medical imaging*, pages 379–387. Springer, 2017.

[Selver *et al.*, 2019] Alper Selver, Ali Emre Kavur, et al. Chaos-combined (ct-mr) healthy abdominal organ segmentation. In *The IEEE International Symposium on Biomedical Imaging (ISBI)*, 2019.

[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.*, 2017] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017.

[Zhang *et al.*, 2018a] Yiheng Zhang, Zhaofan Qiu, Ting Yao, Dong Liu, and Tao Mei. Fully convolutional adaptation networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6810–6818, 2018.

[Zhang *et al.*, 2018b] Yue Zhang, Shun Miao, Tommaso Mansi, and Rui Liao. Task driven generative modeling for unsupervised domain adaptation: Application to x-ray image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 599–607. Springer, 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 *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

[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.

[Zou *et al.*, 2021] Danbing Zou, Qikui Zhu, and Pingkun Yan. Unsupervised domain adaptation with dual-scheme fusion network for medical image segmentation. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3291–3298, 2021.