Scientific Research Publishing

Advances in Pure Mathematics, 2021, 11, 628-643
https://www.scirp.org/journal/apm
ISSN Online: 2160-0384
ISSN Print: 2160-0368

Semantic Constraint Based Unsupervised Domain Adaptation for Cardiac Segmentation

Xin Wang[1], Fan Zhu[1*], Yaxin Peng[1], Chaomin Shen[2], Zhen Ye[3], Chaozheng Zhou[3]

[1]College of Science, Shanghai University, Shanghai, China
[2]School of Computer Science and Technology, East China Normal University, Shanghai, China
[3]Shanghai Electric Central Research Institute, Shanghai, China
Email: xinwang@shu.edu.cn, *zhufan5959@shu.edu.cn, yaxin.peng@shu.edu.cn, cmshen@cs.ecnu.edu.cn, zhye1985@aliyun.com, zhouchzh2018@aliyun.com

How to cite this paper: Wang, X., Zhu, F., Peng, Y.X., Shen, C.M., Ye, Z. and Zhou, C.Z. (2021) Semantic Constraint Based Unsupervised Domain Adaptation for Cardiac Segmentation. *Advances in Pure Mathematics*, **11**, 628-643.
https://doi.org/10.4236/apm.2021.116041

Received: May 31, 2021
Accepted: June 27, 2021
Published: June 30, 2021

Abstract

The segmentation of unlabeled medical images is troublesome due to the high cost of annotation, and unsupervised domain adaptation is one solution to this. In this paper, an improved unsupervised domain adaptation method was proposed. The proposed method considered both global alignment and category-wise alignment. First, we aligned the appearance of two domains by image transformation. Second, we aligned the output maps of two domains in a global way. Then, we decomposed the semantic prediction map by category, aligning the prediction maps in a category-wise manner. Finally, we evaluated the proposed method on the 2017 Multi-Modality Whole Heart Segmentation Challenge dataset, and obtained 82.1 on the dice similarity coefficient and 4.6 on the average symmetric surface distance, demonstrating the effectiveness of the combination of global alignment and category-wise alignment.

Keywords

Medical Image Segmentation, Domain Adaptation, Category-Wise Alignment, Cardiac Segmentation

1. Introduction

Medical image segmentation is a basic task of intelligent medical diagnosis, which aims at extracting target regions like organs, tissues or lesions from medical images. In recent years, deep learning has developed fast in the field of medical image segmentation [1] [2] [3], but still remaining some problems to be solved. On the one hand, deep learning needs sufficient annotated data, but the annotation of medical images is highly cost. On the other hand, deep learning

assumes that the test data and training data are of independent identically distribution, while the distributions of medical image modalities vary largely, as shown in **Figure 1**. So, the segmentation for modality with few annotations is troublesome.

Domain adaptation is one commonly used method to this problem. It aims at transferring the knowledge of labeled data to few labeled or unlabeled data, helping to promote their task performance [4]. In domain adaptation, the labeled data are called source domain data, the few or unlabeled data are called target domain data. When there are no labeled data in the target domain, calling it unsupervised domain adaptation. In this paper, we focus on unsupervised domain adaptation.

Aligning the distributions of source and target domain data is a common strategy for unsupervised domain adaptation. When the distributions of data are aligned, the two data can share one same model. The way of aligning distributions can be divided into two categories: global alignment and category-wise alignment.

The global alignment aligns the marginal distributions of two domains and has been implemented in different spaces.

For example, some unsupervised domain adaptation works implement global alignment in the input image space [5] [6] [7] [8] [9], regarding each input image as a whole sample. By aligning the distributions of input images, the appearance gap of two domains can be narrowed.

Some other works implement global alignment in the feature space [10] [11] [12] [13], taking each feature map as a sample. Once the features of two domains follow the same distribution, they can share one classifier.

In addition, some works implement global alignment in the output space, taking every output map as a sample. The alignment of output maps provides a low computation way for feature alignment, which has been widely used in unsupervised domain adaptation segmentation [14] [15] [16].

Also, there are works that combines the above aspects [17]-[25].

The global distribution alignment can effectively align the marginal distributions of data, but lacking of considering the category information within each data, which may cause the misalignment between categories. So, some works additionally consider the category-wise alignment, to further regularize the segmentation results of each category. Currently, category-wise alignment in unsupervised domain adaptation segmentation is mainly implemented at the feature

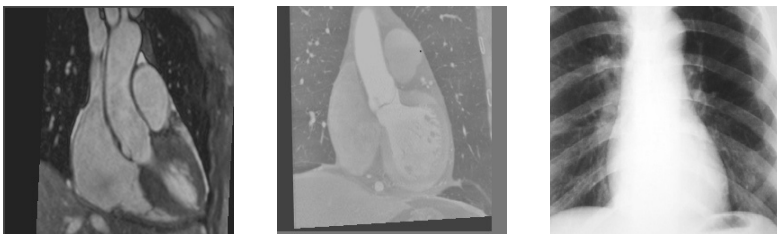


Figure 1. Different modalities of cardiac images.

level and applied in natural image segmentation. For example, Chen *et al.* [26] firstly assign class for each feature vector using the segmentation prediction, then put the features of same category into a discriminator, to align the segmentation results of same category; in a reverse order, Menta *et al.* [27] firstly put the whole feature map into a discriminator, then assign class to the output map of the discriminator; Zhang *et al.* [28] calculate each category's center, and aligns the centers of two domains. In the field of medical image segmentation, category-wise alignment has not been involved yet.

Based on the above works, we propose an improved unsupervised domain adaptation model which combines global alignment and category-wise alignment, and apply it to cross-modality cardiac segmentation. The contributions of the proposed method are as follows:

- Both global distribution alignment and category-wise alignment are introduced to medical image segmentation.
- Category-wise distribution alignment is creatively implemented in the semantic prediction space rather than the feature space.

The organization of the rest of the paper is as follows: in Section 2, we introduce the related works; in Section 3, we illustrate the proposed method; in Section 4, we present and analyze the experimental results; in Section 5, we summarize the whole work.

2. Related Works

2.1. Generative Adversarial Networks

Generative adversarial networks (GAN) [29] is a generative model used to generate data subject to the same distribution of the given data, which is made up of one generator and one discriminator. The generator tries to generate data that looks realistic and the discriminator tries to distinguish between the true and the generated data. So, the two modules form an adversarial relationship. By competing with each other, the two modules mutually promote, finally making the generator generate ideal data. The process of the two modules' competition with each other is called adversarial learning. Due to the unsupervision property of GAN, many unsupervised domain adaptation works [10] [12] [14] [18] use discriminators to align the distributions of two domains.

Image generation is a common application of GAN, but the generation involves randomness, cannot maintaining the image structure. To this problem, Cycle-consistent adversarial networks (CycleGAN) [30] introduces two reversed GANs and a cycle consistency constraint, to maintain the image structure. The first GAN transforms the images to the ideal style, and the second GAN transforms the images of ideal style back to their original style. Thus, after two reversed transformations, the images are reconstructed to their original style. Then, the CycleGAN applies a cycle consistency constraint between the original images and their reconstructed images, promoting the generators to generate structure-invariant images. Many unsupervised domain adaptation works [9] [17] [20] [25] use Cyc-

leGAN to narrow the appearance gap between two domains.

2.2. SIFA

Synergistic Image and Feature Adaptation (SIFA) [20] is an unsupervised domain adaptation method which creatively proposes the synergistic alignments of image and feature and achieves great performance in cross-modality medical image segmentation.

SIFA first uses CycleGAN to narrow the appearance gap between two domains for image adaptation. Then, by sharing the encoder of CycleGAN and the segmentation network, the model has two output spaces, SIFA further aligns the outputs of the two spaces for feature adaptation.

As the CycleGAN and segmentation network share the same encoder, when training, the image adaptation and feature adaptation mutual affect, promoting the synergistic adaptation of image and feature.

In this paper, we adopt the synergistic adaptation strategy of SIFA for global alignment.

3. Proposed Method

Our proposed method considers both global distribution alignment and category-wise alignment for unsupervised domain adaptation. Figure 2 shows an overview of the proposed method. For global distribution alignment, we use the strategy proposed by SIFA [20]; for category-wise alignment, we introduce a new module to the semantic prediction space. The introduction of the proposed method is divided into five sections: image modality transformation, segmentation network, global alignment in image generating space, global alignment in semantic prediction space and category-wise alignment in semantic prediction space.

In Figure 2, the blue arrows represent the source domain data flow and the red arrows represent the target data flow. The fill color of the rectangle represents the modality of images, where blue represents the source domain modality, and red represents the target domain modality; in addition, the color

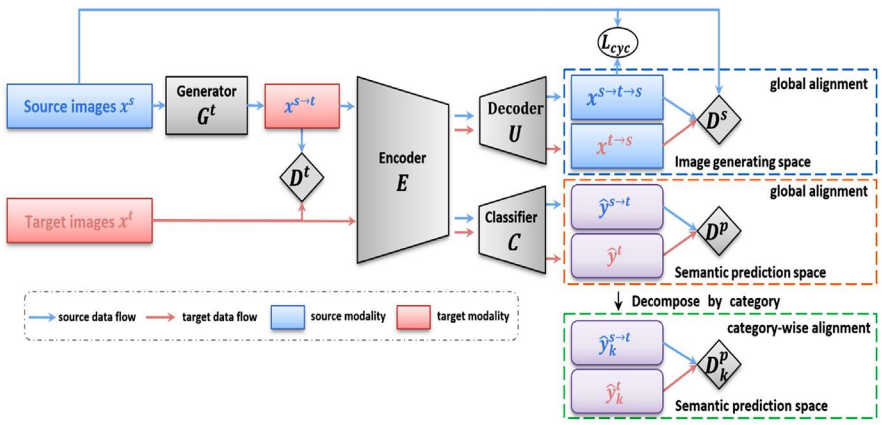


Figure 2. Proposed unsupervised domain adaptation segmentation framework.

of the text also represents the source of the data, blue represents the data come from source domain data x^s, and red represents the data come from target domain data x^t.

3.1. Image Modality Transformation

A large distribution difference exists between cross-modality medical images. If we directly apply the model trained by the source images to the target images, the task performance would be poor. In this section, we use CycleGAN to reduce the appearance gap between two domains.

First, we use a generative adversarial network $\left\{G^t, D^t\right\}$ to transform the modality of the source images to that of the target images. The generator G^t aims to transform the source images x^s into target-like images $G^t\left(x^s\right)=x^{s\to t}$, the discriminator D^t tries to distinguish between the generated target images $x^{s\to t}$ and the real target images x^t. G^t and D^t form a mutual competing relationship. The corresponding objective function of their adversarial learning is:

$$L_{adv}^t\left(G^t, D^t\right)=\mathbb{E}_{x^t\sim X^t}\left[\log D^t\left(x^t\right)\right]+\mathbb{E}_{x^s\sim X^s}\left[\log\left(1-D^t\left(x^{s\to t}\right)\right)\right], \tag{1}$$

where the discriminator D^t aims to differentiate between x^t and $x^{s\to t}$, making $D^t\left(x^t\right)\to 1$ and $D^t\left(x^{s\to t}\right)\to 0$, so D^t aims to maximize the objective function L_{adv}^t; the generator G^t aims to transform x^s into target modality, making $D^t\left(x^{s\to t}\right)\to 1$, so G^t aims to minimize the objective function L_{adv}^t. By the adversarial learning of G^t and D^t, source images are transformed to target-like images.

Then, to preserve the original structure in the transformed images, a reverse generative adversarial network $\left\{E, U, D^s\right\}$ is introduced to transform the target images back to source modality images. Among them, E is an encoder used to map images to a high-dimensional feature space, and U is a decoder mapping the encoded features back to source modality images. $E\circ U$ plays the role of the generator in GAN. Forwarding target images x^t into the generator, it outputs source-like images $x^{t\to s}=U\left(E\left(x^t\right)\right)$, while the discriminator D^s aims to differentiate between the generated source images $x^{t\to s}$ and the real source images x^s. By the adversarial learning of $E\circ U$ and D^s, the target modality images are transformed to source-like images. The objective function of the reverse GAN $\left\{E, U, D^s\right\}$ is:

$$L_{adv}^s\left(E, U, D^s\right)=\mathbb{E}_{x^s\sim X^s}\left[\log D^s\left(x^s\right)\right]+\mathbb{E}_{x^t\sim X^t}\left[\log\left(1-D^s\left(x^{t\to s}\right)\right)\right], \tag{2}$$

where E and U aim to minimize the objective function L_{adv}^s, D^s aims to maximize the objective function.

So the transformations of the two GANs form a cycle, *i.e.*, source images x^s pass through generator G^t and $E\circ U$, obtaining reconstructed source modality images $x^{s\to t\to s}=U\left(E\left(G^t\left(x^s\right)\right)\right)$; target images x^t pass through generator $E\circ U$ and G^t, obtaining reconstructed target modality images

$x^{t\to s\to t} = G^t\left(U\left(E\left(x^t\right)\right)\right)$. By imposing the following cycle consistency constraint to the reconstructed images, generators tend to generate structure-invariant images.

$$L_{cyc}\left(G^t, E, U\right) = \mathbb{E}_{x^s\sim X^s}\left\|x^{s\to t\to s} - x^s\right\| + \mathbb{E}_{x^t\sim X^t}\left\|x^{t\to s\to t} - x^t\right\|. \tag{3}$$

Figure 3 shows the image transformed by the generator G^t. The left is the source image x^s, the medium is the generated image $G^t\left(x^s\right) = x^{s\to t}$, and the right is the target image x^t. It can be seen that the generated image is as the same style as target image while as the same structure as source image.

3.2. Segmentation Network

In Section 3.1, the generator G^t transforms the source images x^s to the target modality images $x^{s\to t}$, thus the transformed images $x^{s\to t}$ and target images x^t are both of target modality, they can share a common segmentation network. As the encoder E learns the features of $x^{s\to t}$ and x^t, we introduce a pixel-wise classifier C after encoder E, forming a segmentation network $E\circ C$.

The training of the segmentation network is supervised by the transformed images $x^{s\to t}$ and their labels. As $x^{s\to t}$ and x^s are of the same structure, they share the same label y^s. Therefore, the objective function of segmentation network is:

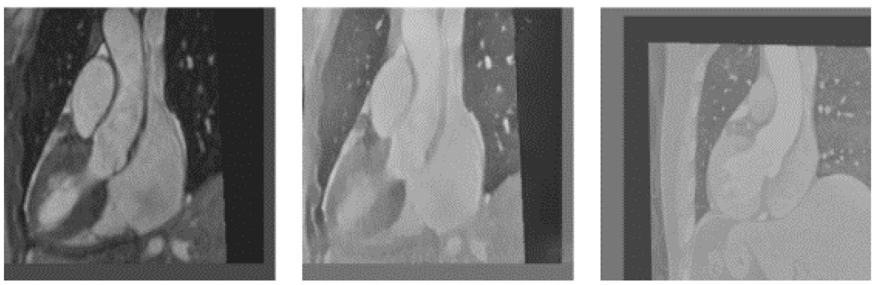
$$L_{seg}\left(E, C\right) = H\left(y^s, \hat{y}^{s\to t}\right) + \alpha\cdot Dice\left(y^s, \hat{y}^{s\to t}\right), \tag{4}$$

where $\hat{y}^{s\to t}$ is the semantic prediction of $x^{s\to t}$, y^s is the one-hot label of $x^{s\to t}$; H is the cross-entropy, Dice is the dice similarity coefficient, α is a hyperparameter using to balance the cross-entropy and the Dice. In the experiment, α is set as 1.

3.3. Global Alignment in Image Generating Space

Image modality transformation alleviates the domain shift between two domains. But when meeting severe domain shift, image adaptation may not be enough to achieve ideal domain adaptation performance. In this section, we further align the distributions of features.

Due to the high dimension of the feature space, aligning features in the feature space takes a lot of computation, so we turn to align the outputs in low-dimensional



x^s $x^{s\to t}$ x^t

Figure 3. Example of image modality transformation.

output space. On the one hand, for domains with the same task, their outputs in low-dimensional space should share a lot of similarity, on the other hand, through back propagation, the distribution of features can be indirectly aligned.

As shown in **Figure 2**, the proposed domain adaptation framework has two output spaces. Call the output space of $E \circ U$ the image generating space, the output space of $E \circ C$ the semantic prediction space. We implement distribution alignment in both these two output spaces. In this section, we first introduce the alignment in image generating space.

The blue dotted box in **Figure 2** corresponds to the image generating space alignment. In image adaptation, discriminator D^s is introduced to differentiate between real and generated source images. For output space alignment, D^s is assigned a new task to distinguish the source of data. Forwarding the generated source modality images $x^{s\to t\to s}$ and $x^{t\to s}$ into discriminator D^s, D^s aims to distinguish the real source of input images, where images $x^{s\to t\to s}$ come from the source domain, images $x^{t\to s}$ come from the target domain. And the corresponding generator $E \circ U$ tries to generate identically distributed images for two domains, confusing the discriminator. When discriminator D^s cannot differentiate the source of two generated images, it demonstrates that the distributions of the two images are quite similar. By the adversarial learning between D^s and $E \circ U$, the distributions of $x^{s\to t\to s}$ and $x^{t\to s}$ are aligned. The corresponding objective function of adversarial learning is:

$$L_{adv}^{\bar{s}}\left(E,U,D^s\right)=\mathbb{E}_{x^{s-t}\sim X^{s\to t}}\left[\log D^s\left(x^{s\to t\to s}\right)\right]+\mathbb{E}_{x^t\sim X^t}\left[\log\left(1-D^s\left(x^{t\to s}\right)\right)\right], \quad (5)$$

where E and U aim to minimize the objective function, and D^s aims to maximize the objective function.

3.4. Global Alignment in Semantic Prediction Space

In this section, we introduce the alignment in the semantic prediction space, corresponding to the orange dotted box in **Figure 2**. Similar to the alignment in the image generating space, a discriminator D^p is introduced to the semantic prediction space to distinguish the source of input data. Forwarding two domains' prediction results $\hat{y}^{s\to t}$ and $\hat{y}^t$ into D^p, D^p aims to distinguish that $\hat{y}^{s\to t}$ comes from the source domain and $\hat{y}^t$ comes from the target domain, while the segmentation network $E \circ C$ tries to align the distributions of these two predictions, confusing the discriminator. At this point, segmentation network $E \circ C$ and discriminator D^p respectively correspond to the role of the generator and the discriminator in GAN. Through the competition between $E \circ C$ and D^p, the two domains' prediction results are aligned. The corresponding objective function is:

$$L_{adv}^{p}\left(E,C,D^p\right)=\mathbb{E}_{x^{s\to t}\sim X^{s\to t}}\left[\log D^p\left(\hat{y}^{s\to t}\right)\right]+\mathbb{E}_{x^t\sim X^t}\left[\log\left(1-D^p\left(\hat{y}^t\right)\right)\right], \quad (6)$$

where $\hat{y}^{s\to t}=C\left(E\left(x^{s\to t}\right)\right)$ is the semantic prediction of the source domain, $\hat{y}^t=C\left(E\left(x^t\right)\right)$ is the semantic prediction of the target domain. E and C aim to

minimize the objective function, D^p aims to maximize the objective function.

Considering that the output space is far away from the shallow layer of the network, the gradient of adversarial learning may not effectively back propagate to the shallow features. Therefore, an auxiliary pixel-wise classifier C_a is further introduced to the second last feature layer, then there appears an additional output space, calling it auxiliary semantic prediction space. Similarly, a discriminator D^{p_a} is introduced to the auxiliary semantic prediction space. By the adversarial learning of $E \circ C_a$ and D^{p_a}, the distributions of the auxiliary semantic predictions are aligned, indirectly aligning the distributions of the shallow features.

The objective function of auxiliary semantic prediction space alignment is:

$$L_{adv}^{p_a}\left(E, C_a, D^{p_a}\right) = \mathbb{E}_{x^{s\to t}\sim X^{s\to t}}\left[\log D^{p_a}\left(\hat{y}_a^{s\to t}\right)\right] + \mathbb{E}_{x^t\sim X^t}\left[\log\left(1 - D^{p_a}\left(\hat{y}_a^t\right)\right)\right]. \quad (7)$$

The segmentation loss function of auxiliary semantic prediction network $E \circ C_a$ is:

$$L_{seg}\left(E, C_a\right) = H\left(y^s, \hat{y}_a^{s\to t}\right) + \alpha \cdot Dice\left(y^s, \hat{y}_a^{s\to t}\right), \quad (8)$$

where $\hat{y}_a^{s\to t} = C_a\left(E\left(x^{s\to t}\right)\right)$ and $\hat{y}_a^t = C_a\left(E\left(x^t\right)\right)$ are the auxiliary semantic prediction results of two domains respectively. And differing from the $E(\cdot)$ in Equation (6), the $E(\cdot)$ in Equation (7) and Equation (8) represents the second last features.

3.5. Category-Wise Alignment in Semantic Prediction Space

The image modality transformation and output space alignment both treat every input or output image as a whole sample, aligning the distributions from a global perspective, without considering the multiple categories within each image. In this section, we introduce category-wise alignment to the proposed framework, further optimizing the alignment between each category.

The proposed category-wise alignment corresponds to the green dotted box in Figure 2. Aligning the features of the same category is a commonly used method for category-wise alignment. The disadvantages of this kind of method is that it requires a large amount of computation due to the high dimension of features and the categories of feature vectors need to be inferenced by the segmentation results, to some extent inconvenient. Intuitively, we can try to implement category-wise alignment directly in the semantic prediction space. On the one hand, for medical image, each category of its foreground corresponds to a structure of human body, whose segmentation result should share a lot of shape and position consistency between domains. On the other hand, the segmentation result directly gives the probability of each pixel belonging to each category. Meanwhile, the dimension of the semantic prediction space is much lower than the feature space, saving a considerable amount of computation. Based on the above points, in this section, we align the category-wise distribution in the semantic prediction space.

The output of the semantic prediction space is a multi-channel segmentation map, each channel corresponding to a category. The value of each pixel in each channel represents the confidence of the pixel belonging to that category. For category-wise alignment, we first decompose the output segmentation map by channel, obtaining the segmentation prediction of each category, as shown in Figure 4. Then, the segmentation maps of the same category are forwarded into a category-dependent discriminator D_k^p, where k represents the category. The discriminator D_k^p aims to differentiate the source of the segmentation maps of the k_{th} category, and the generator $E \circ C$ aims to align the distributions of the two maps, confusing the discriminator. By the adversarial learning of $E \circ C$ and D_k^p, the prediction results of the k_{th} category are aligned.

For medical segmentation result, each category of the foreground corresponds to a specific structure, while the background contains multiple structures. So, we only consider the categories of the foreground when aligning the category-wise distribution. That means, for segmentation task of K categories, there are $K-1$ discriminators introduced to our category-wise alignment. The objective function of the adversarial learning between $E \circ C$ and D_k^p is:

$$L_{adv}^{p(k)}\left(E,C,D_k^p\right)=\mathbb{E}_{x^{s\to t}\sim X^{s\to t}}\left[\log D_k^p\left(\hat{y}_k^{s\to t}\right)\right]+\mathbb{E}_{x^t\sim X^t}\left[\log\left(1-D_k^p\left(\hat{y}_k^t\right)\right)\right], \quad (9)$$

where $k\in\{1,2,\cdots,K-1\}$ represents the category, $\hat{y}_k^{s\to t}=C\left(E\left(x^{s\to t}\right)\right)_k$ and $\hat{y}_k^t=C\left(E\left(x^t\right)\right)_k$ are respectively the segmentation results of the k_{th} category of two domains. E and C aim to minimize the objective function, D_k^p aims to maximize the objective function.

Overall, the full objective function of the proposed method is:

$$\begin{aligned}
&L\left(G^t,D^t,E,U,D^s,C,C_a,D^p,D^{p_a},D_k^p\right)\\
&=\lambda_{adv}^t L_{adv}^t\left(G^t,D^t\right)+\lambda_{adv}^s L_{adv}^s\left(E,U,D^s\right)+\lambda_{cyc}L_{cyc}\left(G^t,E,U\right)\\
&\quad+\lambda_{seg}^1 L_{seg}\left(E,C\right)+\lambda_{seg}^2 L_{seg}\left(E,C_a\right)+\lambda_{adv}^{\bar{s}} L_{adv}^{\bar{s}}\left(E,U,D^s\right)\\
&\quad+\lambda_{adv}^p L_{adv}^p\left(E,C,D^p\right)+\lambda_{adv}^{p_a} L_{adv}^{p_a}\left(E,C_a,D^{p_a}\right)\\
&\quad+\sum_{k=1}^{K-1}\lambda_{adv}^{p(k)} L_{adv}^{p(k)}\left(E,C,D_k^p\right),
\end{aligned} \quad (10)$$

where all the discriminators aim to maximize the above objective function, other modules aim to minimize the objective function. All the modules update in an alternative way. The parameters

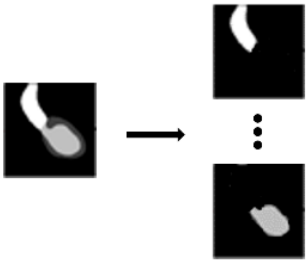


Figure 4. Decomposition of the segmentation map by category.

$$\left\{\lambda_{adv}^{t}, \lambda_{adv}^{s}, \lambda_{cyc}, \lambda_{seg}^{1}, \lambda_{seg}^{2}, \lambda_{adv}^{\bar{s}}, \lambda_{adv}^{p}, \lambda_{adv}^{p_a}, \lambda_{adv}^{p(k)}\right\}$$

are used to balance the functions in the full objective function, setting as

$$\{0.1, 0.1, 10, 1, 0.1, 0.1, 0.1, 0.01, 0.005\}.$$

3.6. Network Configuration and Training

This section introduces the configuration and training details of the proposed network.

We divide all the modules into five categories: generator G^t, encoder E, decoder U, pixel-wise classifier C and C_a, and discriminators $D^t, D^s, D^p, D^{p_a}, D_k^p$. Following the settings in SIFA [20], the generator contains three convolution layers, nine residual blocks, two deconvolution layers and two convolution layers in turn; the encoder contains three convolution-pooling operations, eight residual blocks, two deconvolution layers and two convolution layers; the decoder consists of one convolution layer, four residual blocks, three deconvolution layers and one convolution layer; the classifiers contain a convolution layer and an up-sampling operation; and all the discriminators consist of five convolution layers.

When training, the batch size is set as 8, the learning rate is set as 2×10^{-4}, and the optimizer is set as Adam optimizer. The alternate update order of modules is:

$$G^t \to D^t \to E, C, C_a \to U \to D^s \to D^p \to D_k^p \to D^{p_a} \to G^t$$

4. Experimental Results

4.1. Dataset

The proposed method is evaluated on the cardiac dataset of Multi-Modality Whole Heart Segmentation Challenge 2017 (MMWHS2017) [31]. The dataset contains 20 MRI and 20 CT volumes, which are unpair and from different patients. Four important cardiac substructures not covering each other in 2D coronal view are selected for segmentation, respectively, the ascending aorta (AA), the left atrium blood cavity (LAC), the left ventricle blood cavity (LVC) and the myocardium of the left ventricle (MYO). The adaptation direction is from MRI to CT, which means that MRI is the source, CT is the target.

For MRI, sixteen volumes are used for training and four for validation. For CT, fourteen volumes are used for training, two volumes for validation and others for testing. When training, we use the processed coronal slices provided by SIFA [20], which are clipped, resampled, standardized and enhanced on the basis of the original MMWHS2017 dataset.

4.2. Evaluation Metrics

We use two commonly used metrics in segmentation for evaluation, which are dice similarity coefficient (Dice) and average symmetric surface distance (ASSD).

Dice measures the volume overlap between the predicted result and the ground truth, and ASSD measures the surface distance between this two. The higher Dice and lower ASSD represents the better prediction result. The expressions of Dice and ASSD are as follows:

$$\text{Dice} = \frac{2(A \cap B)}{A + B} \tag{11}$$

$$\text{ASSD} = \frac{\left| \sum_{a \in S(A)} \min_{b \in S(B)} \|a - b\| + \sum_{b \in S(B)} \min_{a \in S(A)} \|b - a\| \right|}{S(A) + S(B)}. \tag{12}$$

where A and B represent the 3D prediction result and the ground truth respectively, $S(\cdot)$ represents the set of voxels in 3D surface.

4.3. Numerical Results

Table 1 shows the numerical results of different methods for cardiac CT substructure segmentation. The left column lists the methods for comparison, in order: without domain adaptation, SIFA, our proposed method and CT supervision. Among them, without domain adaptation (w/o adaptation) means directly applying the segmentation model trained by MRI images to test CT images without using any domain adaptation method. CT supervision means using the segmentation model trained by labeled CT images for CT testing. These two methods respectively provide the lower and higher bound for unsupervised domain adaptation. For fairly comparison, the two methods use the segmentation branch $E \circ C$ in Figure 2 as their segmentation model.

Comparing the results of w/o adaptation and two domain adaptation methods, in both cases of no labeled images in target domain, the domain adaptation methods increase the Dice from 26.7 to 80.0 and 82.1, and decrease the ASSD from 24.5 to 6.0 and 4.6, which greatly improves the numerical performance, demonstrating the effectiveness of domain adaptation.

In addition, comparing our proposed method with SIFA. Our method additionally considers category-wise alignment than SIFA, which increases Dice by

Table 1. Numerical results of different methods on CT cardiac substructure segmentation.

Methods	Dice					ASSD				
	AA	LAC	LVC	MYO	Mean	AA	LAC	LVC	MYO	Mean
W/o adaptation	29.4 (17.9)	53.5 (30.9)	4.7 (7.1)	19.0 (14.1)	26.7	36.2 (31.5)	11.8 (4.5)	29.7 (17.1)	20.5 (8.1)	24.5
SIFA [20]	87.6 (4.2)	86.7 (5.0)	80.0 (6.8)	65.8 (8.9)	80.0	5.6 (3.0)	4.1 (2.3)	7.0 (5.6)	7.3 (4.0)	6.0
Ours	**89.0 (3.1)**	**88.0 (2.9)**	**82.8 (5.3)**	**68.6 (5.6)**	**82.1**	**3.1 (0.7)**	**3.8 (2.5)**	**4.1 (1.3)**	**7.2 (4.3)**	**4.6**
CT supervision	81.7 (24.4)	90.1 3.0)	92.2 (2.0)	87.0 (2.6)	87.7	2.7 (2.2)	2.8 (1.7)	1.6 (0.3)	1.9 (0.5)	2.2

2.1 and decreases ASSD by 1.4, further reducing the numerical difference between the unsupervised domain adaptation methods and the CT supervision. Moreover, it can be seen in Table 1 that, the numerical results for each substructure are all improved, verifying the validity of the category-wise alignment.

4.4. Visualization Results

We represent the visualized segmentation results of different methods in Figure 5. From left to right are respectively: CT test image, w/o adaptation, SIFA, our method, CT supervision, and ground truth. The correspondence between color and cardiac substructure is shown in the right legend.

As shown in Figure 5, the segmentation results of w/o adaptation are messy and irregular. After using domain adaptation, the segmentation shape of each substructure become clear, demonstrating the effectiveness of domain adaptation. Then, comparing the segmentation results between domain adaptation methods and CT supervision, it shows that the performance of the domain adaptation methods is very close to that of the CT supervision, even in the cases of no labeling in the target domain.

Then, we compare the performance of our proposed method and SIFA row-by-row. The first and second rows show that SIFA under segments and over segments some substructures, for example, AA is omitted in the first row and part of the background is mistakenly divided into AA in the second row. Our proposed method improves these deficiencies. In the third row, the segmented LAC by SIFA is not completely closed, existing a small hole, and in the fourth row, the shape of MYO is discontinuous and the segmentation of LVC is also inaccurate. Our method improves the segmentation continuity of substructures and the cohesion between substructures.

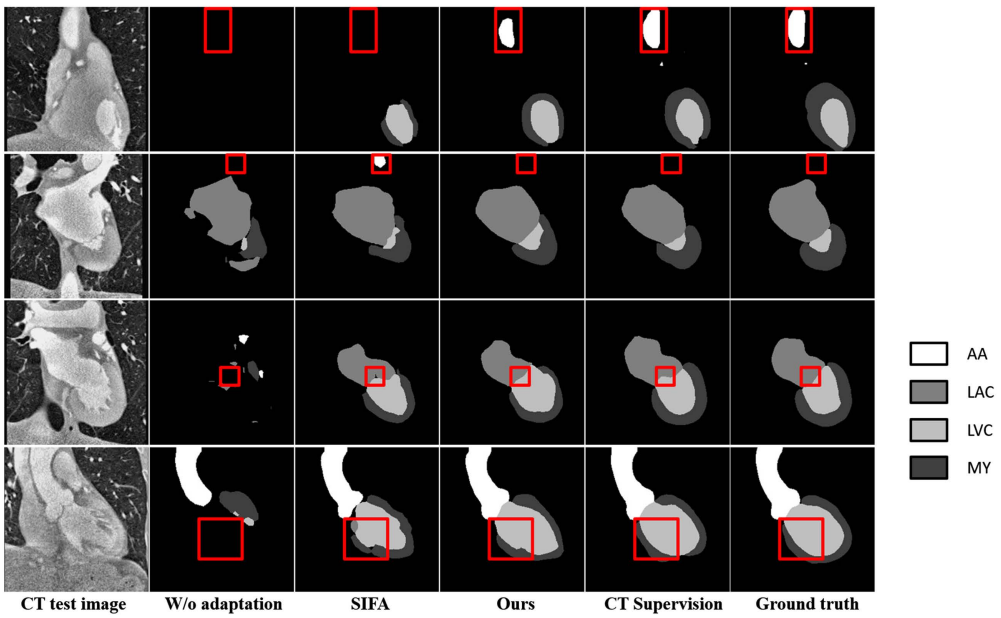


Figure 5. Visualized segmentation results of different methods on CT test data.

5. Conclusions

This paper proposes an improved method for unsupervised domain adaptation segmentation. On the basis of domain adaptation algorithm SIFA, the proposed method further introduces category-wise alignment to the semantic prediction space. Thus, our proposed method both considers global distribution alignment and category-wise distribution alignment. The overall work mainly includes three parts: image modality transformation, global alignment in two output spaces, and category-wise alignment in the semantic prediction space. First, we transform the modality of source images into target in a structure-preserve manner, reducing the distribution difference between two domains in the image level. Then, we respectively introduce discriminators into the image generation space and the semantic prediction space, by aligning the distributions of two domains' outputs, the domain shift is further alleviated. Finally, by equipping each category a discriminator, we align the semantic prediction results in a category-wise manner, further improving the performance of unsupervised domain adaptation. The proposed method is evaluated on the MMWHS2017 cardiac dataset in a direction of MRI to CT. The experimental results show that the proposed method improves the performance of unsupervised domain adaptation.

For future research, we consider introducing the category-wise alignment into the appearance transformation, as the appearance difference of two domains may vary with the region, *i.e.*, for some categories, the difference may be large, while the others may be slight. Taking the category information into consideration may further improve the performance of image appearance transformation.

Funding

This work was supported by the National Natural Science Foundation of China (NSFC, 11771276), and the Shanghai Science and Technology Innovation Action Plan (18441909000).

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Ronneberger, O., Fischer, P. and Brox, T. (2015) U-net: Convolutional Networks for Biomedical Image Segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, Cham, 234-241. https://doi.org/10.1007/978-3-319-24574-4_28

[2] Isensee, F., Jäger, P.F., Kohl, S.A., Petersen, J. and Maier-Hein, K.H. (2019) Automated Design of Deep Learning Methods for Biomedical Image Segmentation.

[3] Milletari, F., Navab, N. and Ahmadi, S.A. (2016) V-net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. 2016 *Fourth International Conference on* 3*D Vision* IEEE, Stanford, 25-28 October 2016, 565-571. https://doi.org/10.1109/3DV.2016.79

[4] Pan, S.J. and Yang, Q. (2009) A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, **22**, 1345-1359. https://doi.org/10.1109/TKDE.2009.191

[5] Zhang, Y., Miao, S., Mansi, T. and Liao, R. (2018) 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*, Springer, Cham, 599-607. https://doi.org/10.1007/978-3-030-00934-2_67

[6] Zhao, S., Li, B., Yue, X., Gu, Y., Xu, P., Hu, R., *et al.* (2019) Multi-Source Domain Adaptation for Semantic Segmentation. 33*rd Conference on Neural Information Processing Systems*, Vancouver, 8-14 December 2019, 7287-7300.

[7] Yang, Y. and Soatto, S. (2020) FDA: Fourier Domain Adaptation for Semantic Segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, 13-19 June 2020, 4085-4095. https://doi.org/10.1109/CVPR42600.2020.00414

[8] Russo, P., Carlucci, F.M., Tommasi, T. and Caputo, B. (2018) From Source to Target and Back: Symmetric Bi-Directional Adaptive GAN. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 8099-8108. https://doi.org/10.1109/CVPR.2018.00845

[9] Chen, C., Dou, Q., Chen, H. and Heng, P.A. (2018) Semantic-Aware Generative Adversarial Nets for Unsupervised Domain Adaptation in Chest X-Ray Segmentation. In: *International Workshop on Machine Learning in Medical Imaging*, Springer, Cham, 143-151. https://doi.org/10.1007/978-3-030-00919-9_17

[10] Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., *et al.* (2016) Domain-Adversarial Training of Neural Networks. *The Journal of Machine Learning Research*, **17**, 2096-2030.

[11] Kamnitsas, K., Baumgartner, C., Ledig, C., Newcombe, V., Simpson, J., Kane, A., *et al.* (2017) Unsupervised Domain Adaptation in Brain Lesion Segmentation with Adversarial Networks. In: *International Conference on Information Processing in Medical Imaging*, Springer, Cham, 597-609. https://doi.org/10.1007/978-3-319-59050-9_47

[12] Dou, Q., Ouyang, C., Chen, C., Chen, H. and Heng, P.A. (2018) Unsupervised Cross-Modality Domain Adaptation of Convents for Biomedical Image Segmentations with Adversarial Loss. *Proceedings of the* 27*th International Joint Conference on Artificial Intelligence*, Stockholm, 13-19 July 2018, 691-697. https://doi.org/10.24963/ijcai.2018/96

[13] Tzeng, E., Hoffman, J., Saenko, K. and Darrell, T. (2017) Adversarial Discriminative Domain Adaptation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, 21-26 July 2017, 7167-7176. https://doi.org/10.1109/CVPR.2017.316

[14] Tsai, Y.H., Hung, W.C., Schulter, S., Sohn, K., Yang, M.H. and Chandraker, M. (2018) Learning to Adapt Structured Output Space for Semantic Segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 7472-7481. https://doi.org/10.1109/CVPR.2018.00780

[15] Wang, S., Yu, L., Yang, X., Fu, C.W. and Heng, P.A. (2019) Patch-Based Output Space Adversarial Learning for Joint Optic Disc and Cup Segmentation. *IEEE Transactions on Medical Imaging*, **38**, 2485-2495. https://doi.org/10.1109/TMI.2019.2899910

[16] Wang, S., Yu, L., Li, K., Yang, X., Fu, C.W. and Heng, P.A. (2019) Boundary and Entropy-Driven Adversarial Learning for Fundus Image Segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, Cham, 102-110. https://doi.org/10.1007/978-3-030-32239-7_12

[17] Hoffman, J., Tzeng, E., Park, T., Zhu, J.Y., Isola, P., Saenko, K., *et al.* (2018) Cycada: Cycle-Consistent Adversarial Domain Adaptation. *International Conference on Machine Learning*, Stockholm, 10-15 July 2018, 1989-1998.

[18] Dou, Q., Ouyang, C., Chen, C., Chen, H., Glocker, B., Zhuang, X. and Heng, P.A. (2019) PnP-AdaNet: Plug-and-Play Adversarial Domain Adaptation Network at Unpaired Cross-Modality Cardiac Segmentation. *IEEE Access*, **7**, 99065-99076. https://doi.org/10.1109/ACCESS.2019.2929258

[19] Zhang, Y., Qiu, Z., Yao, T., Liu, D. and Mei, T. (2018) Fully Convolutional Adaptation Networks for Semantic Segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 6810-6818. https://doi.org/10.1109/CVPR.2018.00712

[20] Chen, C., Dou, Q., Chen, H., Qin, J. and Heng, P.A. (2020) Unsupervised Bidirectional Cross-Modality Adaptation via Deeply Synergistic Image and Feature Alignment for Medical Image Segmentation. *IEEE Transactions on Medical Imaging*, **39**, 2494-2505. https://doi.org/10.1109/TMI.2020.2972701

[21] Zhang, Y. and Wang, Z. (2020) Joint Adversarial Learning for Domain Adaptation in Semantic Segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, **34**, 6877-6884. https://doi.org/10.1609/aaai.v34i04.6169

[22] Yang, J., Xu, R., Li, R., Qi, X., Shen, X., Li, G. and Lin, L. (2020) An Adversarial Perturbation-Oriented Domain Adaptation Approach for Semantic Segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, **34**, 12613-12620. https://doi.org/10.1609/aaai.v34i07.6952

[23] Yang, X., Dou, H., Li, R., Wang, X., Bian, C., Li, S., *et al.* (2018) Generalizing Deep Models for Ultrasound Image Segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, Cham, 497-505. https://doi.org/10.1007/978-3-030-00937-3_57

[24] Dong, J., Cong, Y., Sun, G., Zhong, B. and Xu, X. (2020) What Can Be Transferred: Unsupervised Domain Adaptation for Endoscopic Lesions Segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, 13-19 June 2020, 4023-4032. https://doi.org/10.1109/CVPR42600.2020.00408

[25] Chen, C., Dou, Q., Chen, H., Qin, J. and Heng, P.A. (2019) Synergistic Image and Feature Adaptation: Towards Cross-Modality Domain Adaptation for Medical Image Segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, **33**, 865-872. https://doi.org/10.1609/aaai.v33i01.3301865

[26] Chen, Y.H., Chen, W.Y., Chen, Y.T., Tsai, B.C., Frank Wang, Y.C. and Sun, M. (2017) No More Discrimination: Cross City Adaptation of Road Scene Segmenters. *Proceedings of the IEEE International Conference on Computer Vision*, Venice, 22-29 October 2017, 1992-2001. https://doi.org/10.1109/ICCV.2017.220

[27] Menta, M., Romero, A. and van de Weijer, J. (2020) Learning to Adapt Class-Specific Features across Domains for Semantic Segmentation.

[28] Zhang, Q., Zhang, J., Liu, W. and Tao, D. (2019) Category Anchor-Guided Unsupervised Domain Adaptation for Semantic Segmentation. 33*rd Conference on Neural Information Processing Systems*, Vancouver, 8-14 December 2019, 435-445.

[29] Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., *et*

al. (2014) Generative Adversarial Networks. *Proceedings of the* 27*th International Conference on Neural Information Processing Systems*, Volume 2, 2672-2680.

[30] Zhu, J.Y., Park, T., Isola, P. and Efros, A.A. (2017) Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. *Proceedings of the IEEE International Conference on Computer Vision*, Rio de Janeiro, 14-20 October 2007, 2223-2232. https://doi.org/10.1109/ICCV.2017.244

[31] Zhuang, X. and Shen, J. (2016) Multi-Scale Patch and Multi-Modality Atlases for Whole Heart Segmentation of MRI. *Medical Image Analysis*, **31**, 77-87. https://doi.org/10.1016/j.media.2016.02.006