Contrastive Image Synthesis and Self-supervised Feature Adaptation for Cross-Modality Biomedical Image Segmentation

Anonymous ICCV submission

Paper ID ****

Abstract

This work presents a novel framework CISFA *(**C**ontrastive **I**mage synthesis and **S**elf-supervised **F**eature **A**daptation) that builds on image domain translation and unsupervised feature adaptation for cross-modality biomedical image segmentation. Different from existing approaches, our method employs a one-sided generative model and incorporates a weighted patch-wise contrastive loss between sampled patches of the input image and the corresponding synthetic image, which serves as shape constraints. Furthermore, we notice that the generated images and input images share similar structural information but are in different modalities. To address this, we enforce contrastive losses on the generated images and the input images to train the encoder of a segmentation model to minimize the discrepancy between paired images in the learned embedding space. Compared with existing works that rely on adversarial learning for feature adaptation, such a method enables the encoder to learn domain-independent features in a more explicit way. We extensively evaluate our methods on segmentation tasks containing CT and MRI images for abdominal cavities and whole hearts. Experimental results show that the proposed framework not only outputs synthetic images with less distortion of organ shapes, but also outperforms state-of-the-art domain adaptation methods.*

1. Introduction

Due to the nature of supervised learning, the performance of deep neural networks (DNN) suffers from severe degradation when the domain distribution shifts [27, 8, 28]. One common scenario where such shifts occur is among biomedical images. In clinical practice, computed tomography (CT) and magnetic resonance imaging (MRI) are two common medical radiological imaging techniques. The distinct imaging mechanisms result in dissimilarity of CT and MRI with respect to brightness, contrast, and texture. A deep learning model trained on CT images to segment brain tumours may fail to achieve comparable accuracy on a brain MRI scan. Besides modality, the difference between CT scanners made by different manufacturers can also cause the accuracy to decrease. However, it is infeasible to collect labeled datasets for all possible biomedical image domains, as pixel-wise labelling is laborious and requires expert knowledge, not to mention the privacy issues.

All the above obstacles give rise to unsupervised domain adaptation (UDA) study. There have been numerous works aiming at improving semantic segmentation performance on target domain with only the source domain annotated. One branch of these approaches is to train a single model that is capable of segmenting different styled images, and the key is to extract common features shared by different domains through the encoding path [26]. However, this kind of feature adaptation is coarse-grained. Images with different structural information are forced to be similar in the embedding space, and there could still be a distribution margin between two domains that the discriminator fails to detect. In contrast, [31] applied the L2 norm of the difference between features and the category anchors to drive the intra-category features closer regardless of domains. However, they assumed that a model pretrained on the source domain can generate reliable pseudo-label for target domain images, which is not necessarily true especially when the domain shift is dramatic.

Since the unpaired image translation problem was well tackled by CycleGAN[33], taking advantage of generative models has become another stream in the field of UDA segmentation. The widely used strategy is integrating a segmentation model with the CycleGAN framework. After the source domain images are transferred to the target domain, the corresponding labels would supervise the segmentation training in the target domain [13, 14, 3, 25]. One drawback of CycleGAN based approaches is too many models in the overall framework, especially as some works add additional discriminators and encoding models for feature adaptation. The large number of models makes it difficult to optimize the parameters and lengthens training iterations. Moreover,

although there is an identity loss function in CycleGAN that drives the reconstructed image to be exactly the same as the input image, there is no direct constraint on the input image and the translated image to avoid spatial distortion or structure variation.

Being aware of previous works' limitations, this paper proposes a novel framework *CISFA* featuring lighter generative models and self-supervised feature adaptation, comprising of a generator for image synthesis and a segmenter for target domain. For image synthesis, inspired by a new image translation model named CUT[22], we remove the path that translates target domain to source domain in the CycleGAN flow to facilitate training. To exert the shape-consistency constraint, we maximize the mutual information for image patches at the same position between input images and translated images in latent space. We further extend the pixel-wise InfoNCE loss in CUT by introducing an attention mechanism generated from segmentation labels. This is done by adding weights to the contrastive loss for non-background pixels so that the area of interest is emphasized more during translation. On the other hand, ideally the segmenter only sees target domain images, but there inevitably exists gap between the distribution of generated dataset and real objective dataset. Consequently, feature adaptation is still beneficial for the segmenter. Observing that a successful generator outputs an image that has exactly the same contents and details as the input image except for modality, we realize these paired images serve as good examples of a positive pair in [4]. Thus, we decide to make use of self-supervised learning to conduct feature level adaptation for the segmentation model, different from all previous works. In detail, we add a multi-layer perceptron to the encoder of the segmenter, and project input images as well as synthetic images to one-dimensional features. Then the encoder learns domain invariant features by reducing the cosine distance of paired features.

We conduct experiments with our method and the state-of-the-arts on two medical image UDA tasks. For the first task, we collect 20 CT and 30 MRI abdomen scans from two public datasets, and for the second, we use the MWWHS dataset that contains 20 CT and 20 MRI whole heart 3D images. According to the experiment results, *CISFA* demonstrates superior performance in terms of segmentation accuracy on the target domain than existing works.

2. Related Works

Domain Adaption for Semantic Segmentation One straightforward solution to this problem is training the model to learn domain-independent features. [20] added two classifiers to the encoder in order to get diverse view for a feature, and they extended traditional adversarial loss with an adaptive loss, which was calculated by the cosine distance between the output of those two classifiers. Therefore, the follow-up discriminator would focus more on poorly aligned categories. Instead of using adversarial learning, [31] tried to aggregate features belonging to the same categories for different domain images. Since they did not have annotation for the target domain images, they used a model pretrained on the source domain to generate pseudo-labels. [21] fused prior matching to a VAE model to learn a shared feature space between two domains. More specifically, two domains had a shared VAE module and distinct encoder and decoder modules, and they applied adversarial learning to align features extracted from different encoders. The work[29] was also based on VAE, and they drove the distribution of feature maps to the same parameterized variational form.

Image translation is an alternative way to tackle UDA. Compared to feature alignment in latent space, we can just generate images out of an existing dataset that shares a similar distribution as another dataset. After that, we can take advantage of the annotated domain to do supervised training. The feasibility is ascribed to the rapid development of generative adversarial model (GAN), especially the emergence of CycleGAN[33] in unpaired image translation. [13] firstly integrated a segmentation model with the CycleGAN module for road scene images from different sources. At the same time, [14] used the integrated framework to achieve domain adaption on CT-MR abdomen segmentation. SIFA[3] further extended the workflow by making the segmenter and the source domain generators share the encoder, adding a discriminator to the segmentation results as well as the latent embedding in generated image space. Most recently, [25] added an attention mechanism in the form of normalization to features in the generator to improve the quality of synthesized images for SIFA.

Contrastive Learning Recently, contrastive learning, as a self-supervised learning method, gained popularity after the work SimCLR[4], in which, original image and its augmented views are clustered in feature space. This pretext task can learn useful representations that can boost downstream tasks, like image classification and object detection. Follow-up works provided different perspectives of influential factors, such as transformation combinations[5], batch size and momentum encoders[10, 6], as well as supervised class clustering[17]. Due to the powerful ability of representation learning, this emergent technique has been widely applied to medical image domain, including classification[9, 7, 19, 12] and segmentation task[2, 30, 24, 32]. For example, [2, 30] observed the similarity of adjacent slices in 3D medical volumes and explored how this kind of knowledge could improve the segmentation performance with less labeled data. However, there are few works that utilize contrastive loss for unsupervised domain adaptation.

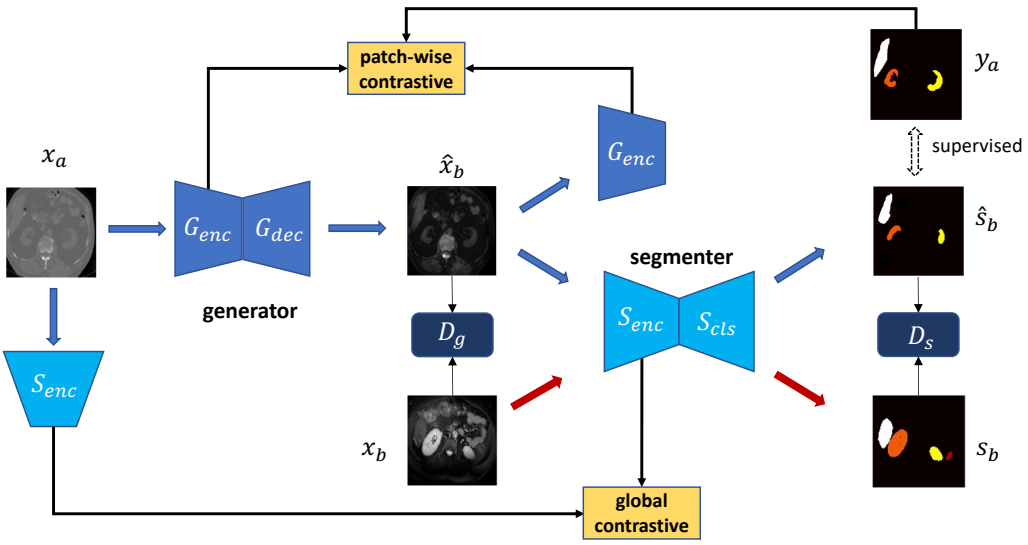

Figure 1. Overview of *CISFA* framework. The main components include a generator G, a segmenter Seg, and two discriminators, D_g & D_s. There are two contrastive losses: a global contrastive loss for self-supervised feature adaptation, and a patch-wise contrastive loss for keeping the shape consistency in synthetic images. Blue/red arrows represent the data flow for source/target domain images respectively.

3. Methods

3.1. Overview

Firstly, we will formulate the UDA segmentation problem for medical imaging. Given two datasets, $A = \{(x_a, y_a)|x_a \in \mathcal{A}\}$ with labels and $B = \{x_b \in \mathcal{B}\}$ without labels, we aim to generate segmentation masks $\hat{y}_b$ for images in B. $\mathcal{A}$ and $\mathcal{B}$ represent the source domain and the target domain respectively, and can be different modalities or be collected from different scanners. Since labels are missing in the target domain, supervised learning is not applicable in this situation. Moreover, as there exists a distribution shift from $\mathcal{A}$ to $\mathcal{B}$, models trained on source dataset fails to give satisfying segmentation accuracy for images in B.

Fig.1 shows the sketch of the overall framework CISFA to tackle the above UDA problem. We use $G(\cdot)$ to translate $x_a \rightarrow \hat{x}_b = G(x_a) \in \hat{\mathcal{B}}$ while maintaining the structural contents in x_a. Then we get labeled data $(\hat{x}_b, y_a)$ to supervise the training of a segmentation model $Seg(\cdot)$. CISFA is different from existing frameworks in four aspects: 1) We use one fewer generative model than CycleGAN based models. This makes the model easier to train. 2) We use two different contrastive losses to deal with structure distortion and domain shift. 3) Our patch-wise contrastive loss, unlike the losses used in the literature for the same purpose, assigns different weights for different patches, enforcing more attention to non-background patches. 4) Unlike existing works that use adversarial learning for feature adaptation, we introduce a global contrastive loss for model to learn domain invariant features. In the rest of this section, we will firstly introduce the detail of the image translation process and then the patch-wise contrastive loss. Subsequently, we will describe the self-supervised feature adaptation for the downsampling path of Seg as well as other losses regarding the segmenter Seg. Lastly, we will present some training details of all models in the framework.

3.2. Weighted Patch-wise Contrastive Loss for Image Synthesis

Similar to all generative adversarial networks, we add a discriminator $D_g(\cdot)$ to distinguish between the target domain image x_a and fake target domain image $\hat{x}_b$. This is done by minimizing the following loss,

$$\begin{aligned} L_{D\_G} =& \mathbb{E}_{x_b \sim B}[log D_g(x_b)] + \\ & \mathbb{E}_{x_a \sim A}[log(1 - D_g(G(x_a))] \end{aligned} \tag{1}$$

Simultaneously, the task of G is to deceive the discriminator into classifying synthetic images as in real $\mathcal{B}$. Gradually, G learns to generate images sharing the similar texture as x_b.

The consistency of structural information between x_a and $\hat{x}_b$ assures that the annotation is still correct for the after-translation image, which is the key to the success of the following supervised training. However, lacking the path that transforms $\hat{x}_b$ back to domain $\mathcal{A}$, *CISFA* has no cycle loss that penalizes the distortion after image translation. Therefore, we devise a patch-wise contrastive loss as an alternative shape consistency constraint. Specifically, we extract feature maps of input images and output images in different levels from the encoder of the generative model G_{enc}, $\{f_a^l | l = 1, 2, ..\} = G_{enc}(x_a)$, $\left\{f_{\hat{b}}^l | l = 1, 2, ..\right\} =$

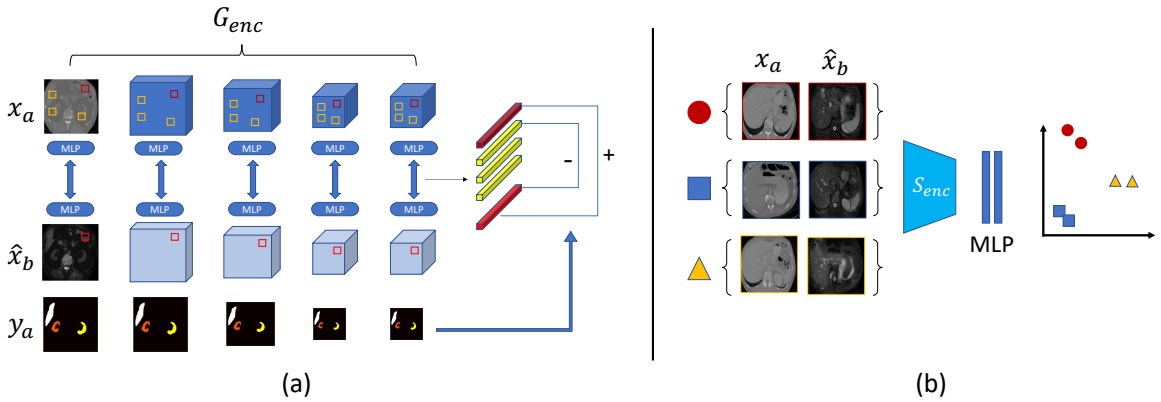

Figure 2. Illustration of the calculation of (a) the patch-wise contrastive loss and (b) global contrastive loss. (a): We select four layers inside the G_{enc} as well as the original images. The small frame in the feature map represents randomly sampled patches. "+" stands for positive pairs and "-" stands for negative pairs. We also downsample the label to the same resolution as each layer of feature maps and increase the weights for non-background patches. (b): Patterns with the same shape and the same color denote positive pairs and should be closer in the latent space.

$G_{enc}(\hat{x}_b)$. For feature map $f_a^l \in \mathbb{R}^{c*h*w}$, where c is the channel number, $h*w$ is the feature map size, l denotes the lth layer. Taking all features into account is computationally expensive. Hence we randomly sample features in dimension $h*w$ for f_a^l and do sampling from the same position for $f_{\hat{a}}^l$. Afterwards, we add a multi-layer perceptron (MLP) as projection head to the sampled features and apply $L2$ normalization. Then we get a feature vector set $F^l = \left\{\mathbf{u}|\mathbf{u} \in \mathbb{R}^{c'}\right\}$, where c' is the feature dimension and is set as 128. Let $\mathbf{u}^+ \in F^l$ denote the feature vector derived from the same position as u in the other feature map. The patch-wise contrastive loss for layer l is defined as

$$L_{pcl}(l,u) = -log\frac{exp(\mathbf{u}\cdot\mathbf{u}^+/\tau)}{\sum_{v\in F^l-u} exp(\mathbf{u}\cdot\mathbf{v}/\tau)} \tag{2}$$

in which, τ is the temperature parameter for the cosine difference and we choose 0.2 for τ in this work. This loss reduces the discrepancy between the same patch before and after transformation in the latent space.

Being aware of the fact that the label y_a is also provided, in this work, we extend the definition of patch-wise contrastive loss by adding weights to features corresponding to different categories. Since the translation correctness of non-background areas is more important to the consequential segmentation training compared to distortion of background pixels, we set the weights of non-background features to be larger than background ones. In practice, for each layer l, we downsample the label to the same resolution as the feature map f^l and decide the weight for each sampled feature vector. The newly added weight $w_p(u)$ is

$$w_p(u) = \begin{cases} 1, & \text{if u is background} \\ w, & \text{else} \end{cases} \tag{3}$$

where w can be any number larger than 1 and we set it as 2 in this work. Then the total loss for generator G is,

$$\begin{aligned} L_G = &\mathbb{E}_{x_a\sim A}[logD_g(G(x_a))] + \\ &\mathbb{E}_{x_a\sim A, l\sim L, u\ \sim F_l}[w_p(u) * L_{pcl}(l,u)] \end{aligned} \tag{4}$$

in which L is the set of selected layers to calculate patch-wise contrastive loss.

3.3. Self-supervised Feature Adaptation

After obtaining the translated image $\hat{x}_b$, we pass it into the segmentation model Seg to get a prediction mask $\hat{s}_b$. By optimizing the dice loss of $(\hat{s}_b, y_a)$, we expect the segmenter to learn how to tackle images in the target domain. As shown in previous works [2, 30], pretraining the encoder path of segmentation model to learn mutual information among similar slices benefits the follow-up segmentation task, where similar slices refer to images at very adjacent positions in the volume. For our case, we observe that x_a and $\hat{x}_b$ share the same content but differ in modality, which are good examples of positive pairs in contrastive learning. Let $S_{enc}(\cdot)$ denote the downsampling path of the segmenter. We project t pairs of input images and their corresponding synthetic images into latent space, and add a MLP head for stronger representation ability in the feature space. After normalization, we get features $\left\{z_i = \left\|\text{MLP}(S_{enc}(x^i))\right\| | i = 1, 2, ..., 2t\right\}$, and assume that $j(i)$ is the index of positive pair feature regarding z_i. We then calculate the contrastive loss similar to SimCLR[4].

$$L_{gcl}(t) = -\frac{1}{2t}\sum_i log\frac{exp(z_i\cdot z_{j(i)}/\tau)}{\sum_{k\neq i}^{2t} exp(z_i\cdot z_k/\tau)}, \tag{5}$$

which is defined as global contrastive loss in this paper. In contrast to L_{ncl}, the feature vector z_i contains the global

information of the input image instead of only one patch. Moreover, we also generate the prediction mask $s_b = Seg(x_b)$ in the real target domain, and add a discriminator D_s to identify the output of images in $\mathcal{B}$. In that case, the segmenter is trained to give segmentation result of the same quality for $\hat{x}_b$ and x_b, even though there is still a minor difference in the distribution of $\hat{\mathcal{B}}$ and $\mathcal{B}$. Then the overall loss for the Seg is

$$\begin{aligned} L_{seg} =& \mathbb{E}_{x_a \sim A}[1 - Dice(Seg(G(x_a), y_a)] + \\ & \mathbb{E}_{t \sim T}[L_{gcl}(t)] + \mathbb{E}_{x_a \sim A}[logD_s(\hat{s}_b)] \end{aligned} \tag{6}$$

where T is the set of batches of paired input and synthetic images. The loss that D_s tries to minimize is formulated as

$$\begin{aligned} L_{D\_S} =& \mathbb{E}_{x_b \sim B}[logD_s(s_b)] + \\ & \mathbb{E}_{x_a \sim A}[log(1 - D_s(\hat{s}_b))] \end{aligned} \tag{7}$$

3.4. Training Strategies

Although we already cut one generative path in our framework compared to CycleGAN, there are still four models for training including G, Seg, D_g and D_s. We integrate all of them into one framework seamlessly and can train all the parameters end-to-end. However, this does not mean we simply calculate all the loss functions and do the backpropagation at the same time. At each training iteration, the order of updating weights is actually $G \rightarrow Seg \rightarrow D_g \& D_s$. Notice that, after changing weights of G, we do an inference on G with the latest weights to update $\hat{x}_b$ and then optimize parameters for Seg. Similarly, we get the prediction masks s_b and $\hat{s}_b$ with the fresh Seg before changing parameters for the two discriminators. Therefore, in the next iteration, when calculating the generative loss for G or Seg, the corresponding classifier has seen the new images or segmentation masks, which is then a fair game for the two min-max game players. After training, we then obtain a segmenter that is capable of making pixel-wise prediction on the target domain without a single label.

4. Experiment

4.1. Dataset

Abdominal Dataset This dataset contains 30 volumes of CT scans from the *Multi-Atlas Labeling Beyond the Cranial Vault Challenge* [1] and 20 volumes of T2-SPIR MRI from the *ISBI 2019 CHAOS CHALLENGE* [16]. We choose four organs that are manually annotated on both datasets as the segmentation task, which are liver, right kidney, left kidney, and spleen. After trimming the whole volume to only contain the region of interest (ROI), we reshape all slices in the transverse plane to be unified 196*196 by interpolation.

MMWHS Dataset [34] provides 20 CT and 20 MRI 3D cardiac images from different patients with annotations by expert radiologists for both modalities, and we focus on 4 anatomical structures as the segmentation objects, including ascending aorta, left atrium blood cavity, left ventricle blood cavity, and myocardium. We also do the crop on the 3D volumes and reshape each images in the coronal view into the size of 160*160.

For preprocessing, we do normalization on cropped images for both datasets so that all pixels in a volume are zero mean and unit variance. We split the two datasets into four folds on volume basis for cross-validation, and then decompose 3D volumes to 2D slices in every fold. This ensures that slices from the same scan can only exist in the same folder. In the training only labels in the source domain are used. For example, if MRI is the target domain for an experiment, we use three folds of CT and MRI slices for training and only CT labels are considered. The remaining CT fold is treated as validation set. After training, we evaluate the segmenter on the remaining MRI fold, from which we compute the dice score as well as average symmetric surface distance (ASSD) for each volume, then report the average and the standard deviation for the fourfold runs.

4.2. Settings

The backbone of our generative model is based on ResNet[11]. We firstly use two convolutional layers with stride equal to 2 to downsample the input image, followed by 9 residual blocks. As for the segmenter, we deploy a U-Net[23] with 4 resolution stages. To reduce memory usage, we build a fully convolutional network with 3 layers as our discriminator architecture, same as that described in [15]. We implement the generative models and discriminator with the deep learning package PyTorch. The GPU devices used are two NVIDIA Tesla P100 with 16GB memory each. The optimizers used for updating weights are all based on the Adam[18] algorithm. The learning rate is also the same for all four models, lr=0.0002, $(\beta_1, \beta_2) = (0.5, 0.999)$. The batch size is set as 4, considering the limitation of memory. The training iteration number is 200 because we observe convergence of losses for all models after training for that number of epochs.

4.3. Baselines

To validate the effectiveness of our proposed method, we conduct experiments with state-of-the-art approaches under the same setting as comparison. CUT[22] is a representative for image translation methods, and the segmentation part is not trained at the same time as the generative model. Accordingly, we just use CUT framework to translate images from source domain to target domain, and use the fake target images with source domain labels to supervise the training of the segmentation model. The reason why we don not include the results of CycleGAN is that the baseline methods below already shows their methods have su-

Table 1. Comparison between state-of-the-art methods and the proposed methods w.r.t. segmentation dice scores on abdominal MRI volumes. The translation direction is CT → MRI. The average dice score and corresponding standard deviation over four independent folds are presented for all four organs, including liver, LK (left kidney, RK(right kidney) and spleen.

Methods	Dice% ↑				
	liver	LK	RK	spleen	avg
Supervised	89.00±1.08	87.19±2.49	83.31±5.05	88.08±1.82	86.90±2.19
W/o adaptation	10.15±3.94	3.67±3.57	4.04±2.95	7.15±6.81	6.25±1.26
CUT[22]	38.17±6.33	32.20±10.69	34.01±9.32	35.83±10.44	35.05±8.19
VarDA[29]	41.63±1.77	32.95±6.47	34.53±4.14	32.23±4.72	35.33±2.60
SASAN[25]	67.23±9.98	61.41±12.95	67.94±14.63	62.63±13.65	64.80±11.48
SIFA[3]	77.24±2.03	68.03±5.60	68.99±5.16	66.79±4.87	70.26±3.69
CISFA(no weight)	76.14±10.72	72.12±4.52	**74.94±4.14**	73.18±3.11	74.10±1.84
CISFA	**80.13±2.21**	**74.45±5.67**	74.51±5.16	**75.86±5.28**	**76.24±2.17**

Table 2. Comparison between state-of-the-art methods and the proposed methods w.r.t. segmentation dice scores on abdominal CT volumes. The translation direction is MRI → CT.

Methods	Dice% ↑				
	liver	LK	RK	spleen	avg
Supervised	89.03±.95	85.53±12.79	83.94±9.46	85.49±4.05	86.00±3.67
W/o adaptation	9.38±3.08	8.88±1.26	8.40±1.31	9.70±1.52	9.09±0.68
CUT[22]	17.78±8.74	28.34±8.05	21.16±11.83	19.29±10.60	21.64±8.66
VarDA[29]	32.78±2.29	38.11±4.17	31.71±4.32	30.26±3.33	33.22±2.38
SASAN[25]	75.36±4.24	67.33±6.43	67.25±6.08	58.70±15.24	67.13±4.32
SIFA[3]	74.03±1.13	65.21±9.88	63.17±10.91	63.53±11.85	66.49±5.61
CISFA(no weight)	**77.45±2.15**	66.91±7.16	64.92±4.57	65.40±13.12	68.67±2.03
CISFA	75.78±3.70	**69.30±7.77**	**70.15±4.77**	**66.57±12.40**	**70.45±2.81**

perior peformance than CycleGAN, but none of them gives the comparison with CUT. VarDA [29] is the latest work that only uses feature adaptation in the field of biomedical UDA segmentation, and its feature adaptation is based on adversarial learning. SIFA [3] and SASAN [25] are the state-of-the-art methods based on synthetic images, derived from CycleGAN framework. For all these baselines, we directly use the codes provided by the authors on github, and the exact same setting is used when comparing these methods with ours, to make a fair comparison. *CISFA* (no weight) and *CISFA* are both our approaches, no weight referring to no weights on L_{pcl}, thus having no information in the label space. Meanwhile, we also provide the segmentation performance of supervised training with all labels on the target domain. We do not need to beat this baseline as it is fully supervised, but can view it as an important reference as the performance ceiling of any UDA methods without labels. On the other hand, w/o adaptation refers to directly applying the model trained on the source domain to target images, which serves as the performance floor.

4.4. Abdominal Image Domain Adaptation

4.4.1 Comparison with the State of the Art

We switch the source and target domains for the bidirectional experiments, and Table 1 present dice score CT → MRI while Table 2 shows MRI → CT results. We also plot the prediction masks of different methods in Fig.3 to visualize segmentation performance. The color coding for different organs in the label space is that white, yellow, orange, and red represents liver, left kidney(LK), right kidney(RK), and spleen, respectively.

We can make several important observations from quantitative and qualitative comparisons. First of all, although there is still a gap compared with the fully supervised method, considering that we do not any target domain annotations, the gap might be reduced if using a very small amount of sparsely labeled data to fine-tune the weights in the segmenter. Thus, our proposed method still proves to be of high clinical practical values. In addition, *CISFA* has the highest dice score for all four organs, and increases the average dice of existing best methods by 5.98 and 2.72 percent in the two tasks. In terms of statistically significance, the p-values for t-test comparing CISFA with CISFA w/o weights, SIFA, and SASAN are 0.0018, < 0.001 and < 0.001, respectively, while in Table. 2, these p-values are 0.015, < 0.001, and < 0.001, respectively. Secondly, VarDA fails to get segmentation performance comparable to other image synthesis approaches in the abdomen dataset. There is a noticeable amount of false positives for liver and spleen in the visual results, which imply that only feature adaptation is not sufficient for significant domain shift cases. Thirdly, in terms of our methods, *CISFA* achieves higher average dice in segmenting target domain images than *CISFA* (no

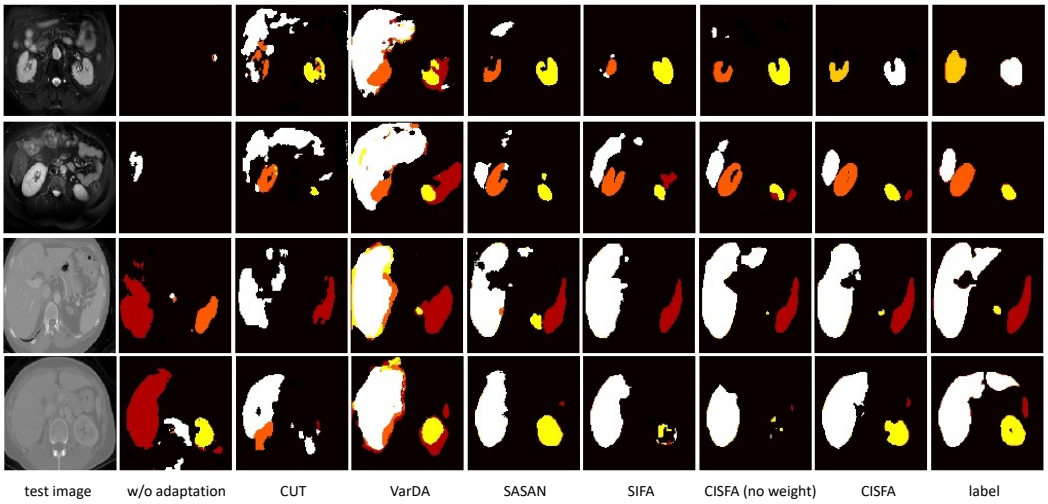

Figure 3. Visualization of segmentation results on the target domain regarding various methods. For the first two rows , the domain translation direction is CT → MRI, and for the other two rows, the direction is MRI → CT.

Table 3. Ablation study of two contrastive losses in CISFA on the CT → MRI adaptation task. Patch-wise contrastive is weighted.

L_{pcl}	L_{gcl}	how L_{gcl} is combined	Dice↑	ASSD↓
		-	39.64	7.63
✓		-	55.76	3.39
	✓	sum	46.63	4.86
✓	✓	sequential	66.50	3.47
✓	✓	sum	**76.24**	**2.52**

weight), which shows the benefits brought by the improved L_{pcl}. Lastly, as shown in Fig.3, without any adaptation technique, the prediction mask is meaningless, while our methods output segmentations that are closest to the ground truth among all the UDA methods. All these illustrate that *CISFA* is an effective complement to current UDA methods in medical imaging segmentation when it is difficult to get target domain annotations.

4.4.2 Ablation Study

Among the contributions of this work, the benefits of introducing weights in patch-wise contrastive loss is clear from the comparison between *CISFA* (no weight) and *CISFA*. The benefits of using contrastive loss over adversarial learning for feature adaptation is clear from the comparison between VarDA and *CISFA*. Here, we further show that both the weighted patch-wise loss L_{pcl} and global contrastive loss L_{gcl} are important in *CISFA* by comparing the dice score and ASSD from different configurations as shown in Table 3. According to the table, removing either of the two losses leads to performance degradation. When no contrastive loss is included, *CISFA* has the lowest segmentation accuracy. On the other hand, the influence of L_{pcl} seems to be more significant than that of L_{gcl}. In our opinion, this is because L_{pcl} directly determines translated image quality, and cutting L_{pcl} means no shape consistency constraint. If organ structures get distorted, it makes no sense for the global contrastive loss to draw images with different content closer in the feature space. Notice that *CISFA* without either component is different from CUT [22]. The segmenter is integrated into the image synthesis flow, and there is G_s to distinguish between prediction masks of real and fake target domain images.

We also explore the impact of how L_{gcl} is combined with other losses in the overall workflow. There are actually two choices: one is to directly add the loss to all the other losses relevant to the segmenter, denoted as "sum" in the table; and the other is to update the weights of the encoder before optimizing other losses for the segmenter at every training iteration, denoted as "sequential". It turns out that the former is better for medical imaging domain adaptation as shown in the table. Although the logic of "sequential" is similar to pretraining in most self-supervised works, the discrepancy in the objectives of sequential weight update process may create problems for the training. After updating the weights in the encoder for minimizing L_{gvl}, the subsequent segmentation training also modifies them but with the purpose of reducing dice loss.

4.4.3 Image Translation

Fig.4 displays original translated images for the two different domain adaptation tasks. In general, all the methods succeed in translating the source domain images to a fake target domain that is quite similar to the real target domain, with only slight differences in brightness, contrast, and quality. However, with a closer look at the generated images, we can observe some deformations and blurring of

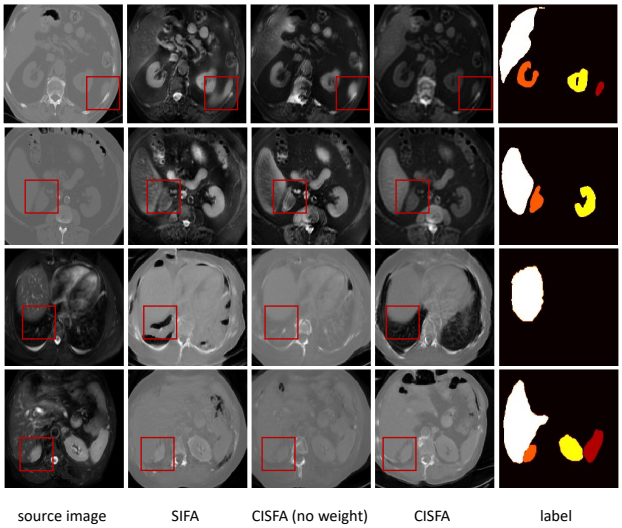

Figure 4. Images of original images and the corresponding synthetic images via SIFA, CISFA (no weight), and CISFA. The upper two rows are translating CT to MRI, and the bottom two rows are translating MRI to CT. We also attach the ground truth of each source image to show the ROI.

the organ regions of interest from SIFA. In the first row, it is evident that the spleen (red) is expanded regarding the shape, and the boundary to background is also blurred for SIFA. Additionally, in the third row SIFA wrongly translates the area in the red box by adding a new structure in the cavity. These distortions are all related to areas containing subject organs, which will then influence the followup supervised segmentation training. When it comes to comparing the two version of our proposed method, in the second and last row, the right kidney and liver areas are more obvious for *CISFA*. This phenomenon is caused by the increased weights on subject organs in our pair-wise contrastive loss, which function as an attention mechanism so that the generator addresses content in the relevant areas more.

4.5. Whole Heart Image Domain Adaptation

We also compare our methods with the state-of-the-art methods under the same setting on MMWHS dataset, as shown in Table 4. It can be noticed that the suprvised method have higher dice score on MMWHS dataset than the abdominal dataset, which is might due to larger number of total slices for all cardiac scans. We can see that if we directly apply the model trained on MRI to CT dataset (w/o adaptation), the dice is lower than that in the abdominal dataset and ASSD could not even be computed because of the large number of false positives, which indicates that domain adaptation is more challenging for this task. The reasons behind the bad performance of "W/o adaptation" have two aspects. On one hand, it reflects that the domain shift between MRI and CT scans in the MMWHS dataset is more drastic than that in the previous abdominal dataset. On the other hand, we observe that there is also a large variation between CT scans and it might due to being collected from different institutions and CT scanners. Therefore, the UDA experiment results have a larger gap to the supervised training performance in MMWHS than the previous dataset. Despite that, we can draw almost the same conclusion from Table 4 as the discussion in section 4.4. Firstly, Feature-alignment methods, like VarDA fails to output a satisfactory segmentation accuracy in contrast to style-transfer methods, like SASAN and SIFA. Secondly, The experiment results show that our method CISFA achieve higher dice score and lowest ASSD than other approaches on this task, which demonstrates that our proposed method can be generalized well to a different dataset. The p-value comparing CISFA with SASAN and SIFA are both less than 0.01.

Table 4. Comparison between the state-of-the-art and the proposed methods on MMWHS dataset for MRI → CT adaptation.

Method	Dice%	ASSD
Supervised	89.78±1.26	0.33±0.05
W/o adaptation	3.13±1.99	-
CUT[22]	37.28±8.32	3.37±1.54
VarDA[29]	40.36±2.86	2.74±0.67
SASAN[25]	61.74±3.34	1.80±0.78
SIFA [3]	64.50±4.21	2.14±1.21
CISFA (ours)	**68.87±3.15**	**1.49±0.31**

5. Conclusion

In this paper, we proposed a novel framework which builds on image domain translation and unsupervised feature adaptation for cross-modality biomedical image segmentation. We introduce a new weighted patch-wise contrastive loss to directly exert shape constraint on the input images and translated images, with special attention on non-background patches. Meanwhile, we innovatively use self-supervised representation learning as feature adaptation to improve segmentation performance. Experiments on two public datasets convinced the superiority of our method over state-of-the-art. In the future, we will further reduce the gap between our methods and the supervised training baselines in order to make our CISFA framework applicable to real clinical scenarios. We might introduce a few sparsely labeled target domain images and test how much we can reduce the annotation efforts if we would like to get the equally accurate segmentation as fully supervised training, which is definitely beyond the scope of unsupervised domain adaptation.

References

[1] "multi-atlas labeling beyond the cranial vault. https://www.synapse.org/#!Synapse:syn3193805/wiki/89480. Accessed: 2021-09-06.

[2] Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu. Contrastive learning of global and local features for medical image segmentation with limited annotations. *arXiv preprint arXiv:2006.10511*, 2020.

[3] 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.

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[5] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*, 2020.

[6] 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.

[7] Ozan Ciga, Tony Xu, and Anne Louise Martel. Self supervised contrastive learning for digital histopathology. *Machine Learning with Applications*, 7:100198, 2022.

[8] Gabriela Csurka. Domain adaptation for visual applications: A comprehensive survey. *arXiv preprint arXiv:1702.05374*, 2017.

[9] Matej Gazda, Ján Plavka, Jakub Gazda, and Peter Drotar. Self-supervised deep convolutional neural network for chest x-ray classification. *IEEE Access*, 9:151972–151982, 2021.

[10] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

[11] 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.

[12] Xuehai He, Xingyi Yang, Shanghang Zhang, Jinyu Zhao, Yichen Zhang, Eric Xing, and Pengtao Xie. Sample-efficient deep learning for covid-19 diagnosis based on ct scans. *medrxiv*, pages 2020–04, 2020.

[13] 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.

[14] Yuankai Huo, Zhoubing Xu, Shunxing Bao, Albert Assad, Richard G Abramson, and Bennett A Landman. Adversarial synthesis learning enables segmentation without target modality ground truth. In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pages 1217–1220. IEEE, 2018.

[15] 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.

[16] A Emre Kavur, N Sinem Gezer, Mustafa Barış, Sinem Aslan, Pierre-Henri Conze, Vladimir Groza, Duc Duy Pham, Soumick Chatterjee, Philipp Ernst, Savaş Özkan, et al. Chaos challenge-combined (ct-mr) healthy abdominal organ segmentation. *Medical Image Analysis*, 69:101950, 2021.

[17] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *arXiv preprint arXiv:2004.11362*, 2020.

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

[19] Bin Li, Yin Li, and Kevin W Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14318–14328, 2021.

[20] Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2507–2516, 2019.

[21] Cheng Ouyang, Konstantinos Kamnitsas, Carlo Biffi, Jinming Duan, and Daniel Rueckert. Data efficient unsupervised domain adaptation for cross-modality image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 669–677. Springer, 2019.

[22] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *European Conference on Computer Vision*, pages 319–345. Springer, 2020.

[23] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[24] Aiham Taleb, Winfried Loetzsch, Noel Danz, Julius Severin, Thomas Gaertner, Benjamin Bergner, and Christoph Lippert. 3d self-supervised methods for medical imaging. *Advances in neural information processing systems*, 33:18158–18172, 2020.

[25] Devavrat Tomar, Manana Lortkipanidze, Guillaume Vray, Behzad Bozorgtabar, and Jean-Philippe Thiran. Self-attentive spatial adaptive normalization for cross-modality domain adaptation. *IEEE Transactions on Medical Imaging*, 2021.

[26] 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.

[27] Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018.

[28] Garrett Wilson and Diane J Cook. A survey of unsupervised deep domain adaptation. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(5):1–46, 2020.

[29] Fuping Wu and Xiahai Zhuang. Unsupervised domain adaptation with variational approximation for cardiac segmentation. *IEEE Transactions on Medical Imaging*, 2021.

[30] Dewen Zeng, Yawen Wu, Xinrong Hu, Xiaowei Xu, Haiyun Yuan, Meiping Huang, Jian Zhuang, Jingtong Hu, and Yiyu Shi. Positional contrastive learning for volumetricmedical image segmentation. *arXiv preprint arXiv:2106.09157*, 2021.

[31] Qiming Zhang, Jing Zhang, Wei Liu, and Dacheng Tao. Category anchor-guided unsupervised domain adaptation for semantic segmentation. *arXiv preprint arXiv:1910.13049*, 2019.

[32] Yejia Zhang, Xinrong Hu, Nishchal Sapkota, Yiyu Shi, and Danny Z Chen. Unsupervised feature clustering improves contrastive representation learning for medical image segmentation. In *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1820–1823. IEEE, 2022.

[33] 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.

[34] 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.