2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 978-1-6654-0126-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/BIBM52615.2021.9669620

AttENT: Domain-Adaptive Medical Image Segmentation via Attention-Aware Translation and Adversarial Entropy Minimization

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

***Abstract*—Due to the intrinsic domain shift among different modalities, it is nontrivial to directly apply a well-trained model into other cross-modality medical images. Unsupervised domain adaptation (UDA) has the potential to reduce such domain shift. However, existing UDA methods try to align domains in either image level or in feature level, failing to consider the unified relationship between cross-modality images and their corresponding features. In this paper, we propose a novel UDA framework for domain adaptive medical image segmentation. The proposed framework synergizes both pixel space and entropy space for domain alignment. Specifically, in the pixel space, we introduce the attention mechanism into CycleGAN, and enhance the semantic and geometric consistency of the target organs during the image style transformation. In entropy space, we utilize entropy minimization principle to force consistent image segmentation between well-annotated source domain and non-annotated target domain. The aligned ensemble of two representation spaces enables a well-trained segmentation model to effectively transfer from source domain to target domain. The experimental results demonstrate the effectiveness of the proposed method. For the task of multi-organs segmentation from cross-modality medical images, our proposed framework achieves state-of-the-art performance, with some specific metric even superior to those of supervised methods. The code is available at https://github.com/lichen14/AttENT.**

***Index Terms*—Deep learning, Unsupervised domain adaptation, Attention-aware image translation, Entropy-based adversarial learning, Abdominal multi-organs segmentation.**

I. Introduction

The training process of supervised networks relies on massive annotated data, which is used as the direct supervision for feature extraction and reconstruction. These supervised mothods have achieved great success in computer vision tasks, such as classification [1]–[3], segmentation [4], [5] and detection [6]–[8]. However, these annotated data help improve the performance while limiting the generalization ability of the networks. In other words, due to the feature distribution gap between datasets from various domains, a fine-grained model trained on well-annotated source dataset can hardly be generalized onto target dataset without significant performance degradation, which is caused by the domain shift generated during the cross-domain transferring. The above phenomenon is particularly noteworthy when applying transfer learning to cross-modality analysis tasks of medical images. Specifically, magnetic resonance imaging (MRI) and computed tomography (CT) are the two most commonly used clinical medical images [9], both of them can be used to scan the abdominal organs. However, these medical images are two domains, whose distribution and representation are quite different but implicit characteristics are shared. Therefore, it is challenging to carry out cross-domain medical image transfer learning between the aforementioned two modalities.

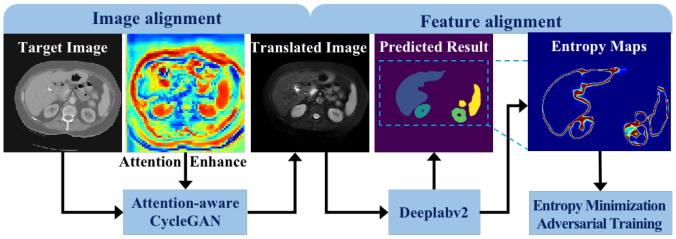


Fig. 1. **Illustration of our method on domain-adaptive medical image segmentation.** The image alignment in pixel space and feature alignment in entropy space are synergized in the proposed AttENT for domain generalization.

In order to solve domain shift in medical image segmentation, we aim to transfer domain-invariable knowledge learned from source domain to target domain based on the idea of unsupervised domain adaptation (UDA) methods. Based on the UDA, we aligned domains on two spaces, i.e., **Att**ention-aware pixel space and **ENT**ropy space. As shown in Fig. 1, we synergized them and proposed a novel medical image segmentation framework called **AttENT**, which can describe the representation of cross-modality image in above spaces and effectively perform unsupervised domain adaptation without any target annotation. Specifically, for the alignment in the pixel space, we introduced the attention mechanism into the CycleGAN [10] with cycle-consistency loss. Besides, the skip connections were bridged between encoder and the decoder in the generator. The attention gates were built on the connections to enhance the learning toward task-related organs. Such design can effectively ensure that the structural information was preserved as much as possible during image style translation. For the alignment in the entropy space,

[†] These authors contributed equally to this work.

978-1-6654-0126-5/21/$31.00 ©2021 IEEE

we redesigned the adversarial learning with the principle of entropy-minimization. By reducing the discrepancy in entropy distribution between the source and target prediction results, domain-adaptive feature alignment can be accomplished. In this way, AttENT can be applied to segmentation of cross-modality medical images and alleviate the annotation workload from supervised learning. Our contributions are as follows:

- We enhance image alignment by introducing the Attention Gates into the CycleGAN architecture, which can effectively preserve the task-related structural information and enhance the learning of target organs. (Shown in Fig.2)
- To the best of our knowledge, we are the first to synergize pixel space alignment and entropy space alignment to characterize domain discrepancy in domain-adaptive medical image segmentation. Fine-grained segmentation across cross-modality datasets gets accomplished by minimizing such entropy. (Shown in Fig.3)
- We evaluate the proposed method on two public datasets (MALBCV ⇔ CHAOS). Our framework is superior to other state-of-the-art UDA methods, with some specific metrics even better than those of supervised method.

II. Related works

The related works can be divided into two mainstreams. The first one is intuitive and straightforward, which is to collect target domain dataset with annotated labels. Then merge the source domain data and target domain data to carry out joint-training [11], semi-supervised learning [12] and fine-tuning [13]. However, such solutions require pixel-wise manual annotation on target images, which is time-consuming and labor-intensive. Therefore, it is not practical for most clinical scenes to provide massive labeled high-quality datasets.

Considering different imaging modalities reflect the same anatomic structures, multi-modality learning has became a promising direction to make up the annotation scarcity in automatic medical image analysis, and has inspired another stream, i.e., unsupervised domain adaptation (UDA) methods. The main goal [14] of UDA is to minimize the discrepancy in feature distribution between the source domain and the target domain, and to learn a cross-domain fine-grained model with strong generalization ability.

Most UDA methods achieve this goal by aligning domains, including image-level alignment [10], [15]–[17] and feature-level alignment [18]–[21]. Image alignment usually utilizes the GANs [22] to convert target images into source-style data. CycleGAN [10] designed generator and discriminator for every domain to get cycle-consistent transformation. Based on CycleGAN, Xue et al. [16] collaborated the edge and mask segmentation.

Feature alignment usually narrows the distribution between domains in the feature space by discriminating the extracted features. The SynSeg-Net [21] and AdaptOut [19] narrowed the distributions gap in the feature space via adaptive adversarial learning. Chen et al. [18] proposed the synergistic alignment from image and feature level. Meanwhile, there is also model alignment [23], [24] carried out in the absence of source domain data. This idea is to generate pseudo-label of target images from source models and train the target model recurrently.

III. Methodology

A. Motivation

Due to the lack of annotated datasets, supervised learning based methods are difficult to be generalized for medical scenarios. In order to solve this problem, we investigated how to train the model directly on unannotated datasets and obtain promising performance. Despite the different representations among images in various modalities, the latent semantic information of organs in these images is the same, i.e., images from different modalities share common semantic feature spaces. Based on the fact, UDA methods are utilized to exploit such shared latent space in medical image segmentation tasks. However, we argued that there were two drawbacks that prevent the existing UDA methods from being directly applied, and our work optimized them in the two spaces.

Firstly, we observed that when the existing UDA methods perform style translation in the pixel space, the structural information of the objects may be lost, thereby reducing the performance of downstream tasks. We analyzed that the loss of the above content is due to the destruction of details caused by the down-sampling in the encoder. Implementing the attention mechanism [25] into generators and enhancing the learning of targets are the best solutions to recover such degradation. After that, attention-aware alignment will improve image translation without structural information loss in the pixel space.

Secondly, although the existing UDA methods can find domain-invariant representations through direct adversarial training in the original feature space, these are lots of information irrelevant to the task, which do not contribute much to the actual downstream task. Inspired by [26], we found that when the source domain model was directly evaluated on target domain, regions with larger errors tend to have higher entropy values. At the same time, for target predicted results with large errors, the entropy map is quite different from the counterpart in the source domain. Based on this observation, following the research of [27], we proposed an entropy-based alignment strategy. This strategy is to reduce the entropy map discrepancy between target domain and the source domain by minimizing the adversarial entropy, and then alleviate domain discrepancy in entropy space.

B. Attention-aware Image Alignment in the Pixel Space

In the background of the unsupervised domain adaptation, images x_s in the source domain $\mathbb{X}_S$ own annotation y_s while images x_t in the target domain $\mathbb{X}_T$ do not. Based on the CycleGAN [10], we introduce the representation map G_{T2S} to translate target images x_t into source style-like images x_{t2s}. In order to protect the structural semantic information in the target image from being destroyed, we refer to the Attention U-Net [25] and introduce Attention Gate into the generators.

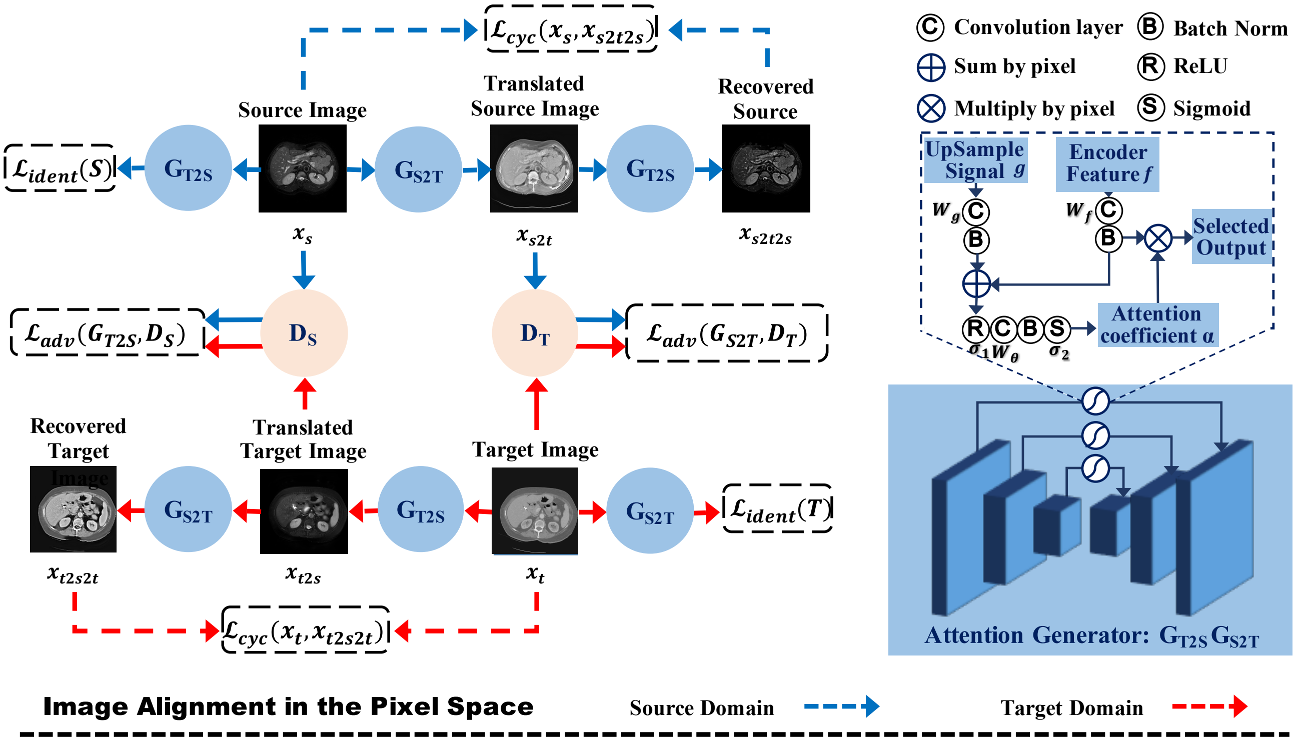


Fig. 2. **Illustration of attention-aware image alignment via redesigned CycleGAN in the pixel space.** The blue arrows and red arrows represent the data flow of source domain (e.g., MRI) and target domain (e.g., CT) respectively. The attention mechanism is introduced into generators and this attention-aware alignment will improve image translation without structural information loss in the pixel space.

Specifically, the AttENT bridges the skip connections between encoder and the decoder. Then attention gates are built on the connections to enhance the learning toward task-related organs. The process of feature selection and the output of Attention Gate (AG) can be formulated as follows:

$$\begin{aligned} AG =& \boldsymbol{f} \times \sigma_2\{W_\theta^T \times \sigma_1 \times \\ & \left[\left(W_f^T \times \boldsymbol{f} + b_f\right) + \left(W_g^T \times \boldsymbol{g} + b_g\right)\right] + b_\theta\}, \end{aligned} \tag{1}$$

where the $\boldsymbol{f}$ is encoder feature while the $\boldsymbol{g}$ means upsample signal. The (W, b) and σ represent the convolution and activation. The structure of the Attention Gate and the process of the image alignment are shown in Fig. 2. Meanwhile, the adversarial discriminator D_S is designed to distinguish the generated source images x_{t2s} and the real source images x_s. The above zero-sum game process can be represented by minimizing the following loss $\mathcal{L}_{adv}(G_{T2S}, D_S, \mathbb{X}_T, \mathbb{X}_S)$:

$$\begin{aligned} \mathcal{L}_{adv} =& \mathbb{E}_{x_s \sim \mathbb{X}_S}\left[\log D_S(x_s)\right] + \\ & \mathbb{E}_{x_t \sim \mathbb{X}_T}\left[\log\left(1 - D_S(G_{T2S}(x_t))\right)\right], \end{aligned} \tag{2}$$

the same is true when converting the source domain to target domain and obtaining the representation map G_{S2T} with the loss $\mathcal{L}_{adv}(G_{S2T}, D_T, \mathbb{X}_S, \mathbb{X}_T)$.

$$\begin{aligned} \mathcal{L}_{adv} =& \mathbb{E}_{x_t \sim \mathbb{X}_T}\left[\log D_T(x_t)\right] + \\ & \mathbb{E}_{x_s \sim \mathbb{X}_S}\left[\log\left(1 - D_T(G_{S2T}(x_s))\right)\right]. \end{aligned} \tag{3}$$

Besides, the cycle consistency loss function $\mathcal{L}_{cyc}(G_{T2S}, G_{S2T}, \mathbb{X}_T, \mathbb{X}_S)$ is also adopted to avoid contradiction between G_{T2S} and G_{S2T} in the conversion process, which is defined as follows.

$$\begin{aligned} \mathcal{L}_{cyc} =& \mathbb{E}_{x_s \sim \mathbb{X}_S}\left[\|G_{T2S}(G_{S2T}(x_s)) - x_s\|_1\right] + \\ & \mathbb{E}_{x_t \sim \mathbb{X}_T}\left[\|G_{S2T}(G_{T2S}(x_t)) - x_t\|_1\right], \end{aligned} \tag{4}$$

we argue that the qualified domain adaptation should have cycle consistency, which means that the reconstructed image from the translated target images should be as consistent as possible with the original images.

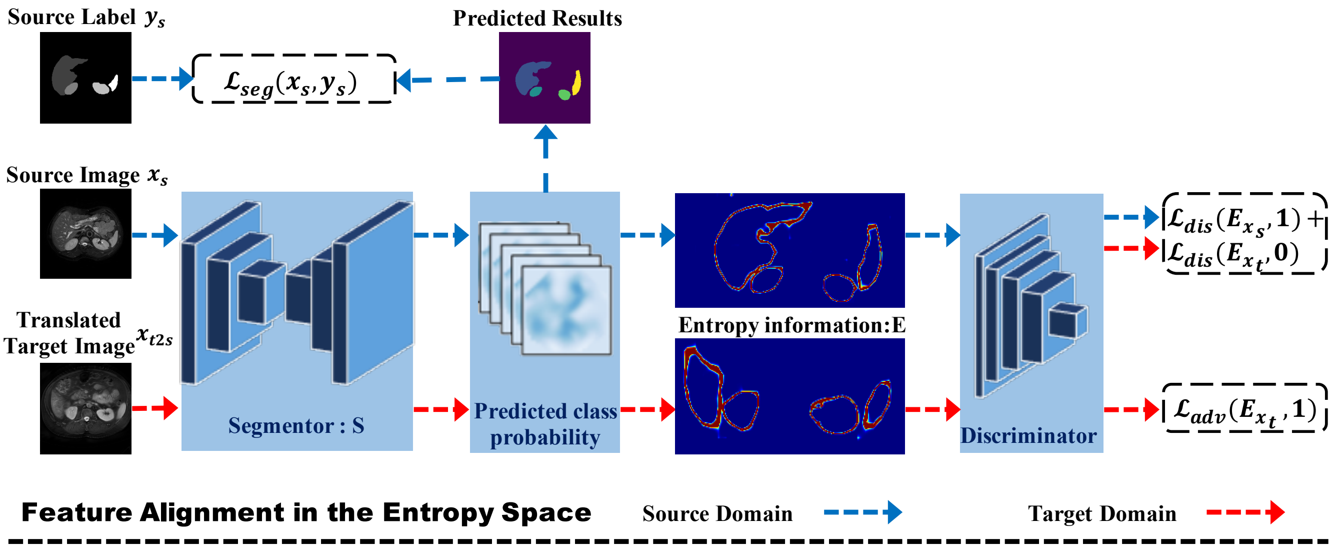
C. Adversarial Feature Alignment in the Entropy Space

The existing entropy loss of the input can be obtained by directly summing the pixel-level normalized entropy, which ignored the relationship between objects in the input. In this section, we adopt the idea in [27] and propose the principle of adversarial entropy minimization to utilize the shared semantic features from cross-domain objects in the entropy space. Firstly, according to the defined segmentation loss $\mathcal{L}_{seg}$, we utilize the labeled source image (x_s, y_s) to train the source domain segmentor S.

$$\mathcal{L}_{seg} = \sum_c^C[-y_s^c \log S(x_s)^c], \tag{5}$$

where C denotes the number of classes. Secondly, we calculate the normalized entropy map E_{x_s} based on the predicted result $S(x_s)$, which is defined as follows:

$$E_{x_s} = -\frac{1}{\log C}\sum_c^C S(x_s)^c \log S(x_s)^c. \tag{6}$$



Source Label y_s
Predicted Results
$\mathcal{L}_{seg}(x_s, y_s)$
Source Image x_s
Translated Target Image x_{t2s}
Segmentor : S
Predicted class probability
Entropy information:E
Discriminator
$\mathcal{L}_{dis}(E_{x_s}, 1) +$
$\mathcal{L}_{dis}(E_{x_t}, 0)$
$\mathcal{L}_{adv}(E_{x_t}, 1)$
Feature Alignment in the Entropy Space
Source Domain
Target Domain

Fig. 3. **Illustration of feature alignment in the entropy space.** The blue arrows and red arrows represent the data flow of source domain and target domain respectively. This workflow is to reduce the entropy map discrepancy between target domain and the source domain by minimizing the adversarial entropy, and then alleviate domain discrepancy in entropy space.

Next, we can get the entropy map E_{x_t} of target image from the segmentor S in the same way. After that, the discriminator D_E is constructed to distinguish entropy map of the source image and the target image. D_E is trained by loss $\mathcal{L}_{dis}$ to minimize the features distribution between source domain and target domain in the entropy space:

$$\mathcal{L}_{dis} = \sum_{x_s}^{\mathbb{X}_S} \mathcal{L}_{D_E}(E_{x_s}, 1) + \sum_{x_t}^{\mathbb{X}_T} \mathcal{L}_{D_E}(E_{x_t}, 0). \quad (7)$$

Finally, we use the adversarial loss $\mathcal{L}_{adv}$ to train segmentor S again and recover the degraded performance due to domain shift, which is formulated as follows:

$$\mathcal{L}_{adv} = \sum_{x_t}^{\mathbb{X}_T} \mathcal{L}_{D_E}(E_{x_t}, 1). \quad (8)$$

The process of the above entropy-based feature alignment is shown in Fig. 3. Different from original alignment in feature space, our method will pay attention to the decrease the discrepancy between the entropy map of source feature and target feature. Because entropy can denote the uncertainty in the matrix while the boundary of object is the most uncertainty pixels in the semantic segmentation task, it is beneficial to apply our entropy-based feature alignment into image segmentation.

IV. Experiments and Results

A. Experimental settings

1) Experimental datasets and preprocessing methods: For the abdominal organs segmentation based on multi-modalities datasets MRI ⇔ CT, we evaluated segmentation performance of the proposed AttENT on four abdominal organs on two public datasets, whose image modalities are MRI and CT.

- CHAOS (Combined Healthy Abdominal Organ Segmentation) [30] is the MRI dataset from the ISBI 2019. We extracted 20 labeled volumes from the T2-SPIR MRI training dataset.
- MALBCV (Multi-Atlas Labeling Beyond the Cranial Vault) is the CT dataset from the MICCAI 2015. We extracted 30 labeled volumes from the CT training dataset.

Although the above datasets have various external appearance, they are both scans of human abdominal organs and manually annotated by experienced specialists. For the common organs in the datasets, including liver, right kidney, left kidney, and spleen, we designed cross-modalities segmentation experiments between these organs.

For the preprocessing methods, according to the settings of SIFA [28], in order to use the 3D volume data to train the 2D segmentation models, we slice the original dataset sequentially along the axial. In the above slicing process, we only selected the slices containing any four target organs, and normalized the 2D slice data with the size 256×256. Finally, we augment the training dataset by resizing and random cropping to reduce overfitting. After that, the datasets are divided into training set and testing set at a ratio of 4:1.

For data augmentation, we only augment the training dataset by resizing and random cropping to reduce overfitting. And the annotated masks of the source domain were only used in the designed domain-adaptive cross-modality segmentation experiment.

2) Evaluation metrics: We adopted the most commonly used metrics in this field for evaluation, i.e., Dice coefficient [31] and Average symmetric surface distance (ASSD) [18].

Dice coefficient is derived from binary classification and is essentially a measure of the overlap between two samples. The indicator range is [0,1], where 1 means complete overlap and

TABLE I

Quantitative results between our AttENT and other SOTA UDA methods for cross-modality abdominal organs segmentation. ASSD and Dice of four organs are compared here, and their average values are calculated. Note that, the smaller for ASSD is better, and reverse on Dice.

MRI ⇒ CT										
Methods	ASSD (voxel)					Dice (%)				
	Liver	Right Kidney	Left Kidney	Spleen	Avg	Liver	Right Kidney	Left Kidney	Spleen	Avg
Supervised	1.0	1.8	0.9	1.2	1.2	92.8	86.4	87.4	88.2	88.7
SynSegNet [21]	2.2	1.3	2.1	2.0	1.9	85.0	82.1	72.7	81.0	80.2
AdaOutput [20]	1.7	1.2	1.8	1.6	1.6	85.4	79.7	79.7	81.7	81.6
CycleGAN [10]	1.8	1.3	1.2	1.9	1.6	83.4	79.3	79.4	77.3	79.9
CyCADA [15]	2.6	1.4	1.3	1.9	1.8	84.5	78.6	80.3	76.9	80.1
SIFA-v1 [18]	2.1	1.1	1.6	1.8	1.6	87.9	83.7	80.1	80.5	83.1
SIFA-v2 [28]	1.2	1.0	1.5	1.6	1.3	88.0	83.3	80.9	82.6	83.7
AttENT	**0.68**	**1.31**	**1.43**	**1.21**	**1.16**	**88.56**	**80.66**	**85.59**	**86.34**	**85.29**
W/o Adapt	2.9	5.6	7.7	7.4	5.9	73.1	47.3	57.3	55.1	58.2

CT ⇒ MRI										
Methods	ASSD (voxel)					Dice (%)				
	Liver	Right Kidney	Left Kidney	Spleen	Avg	Liver	Right Kidney	Left Kidney	Spleen	Avg
Supervised	1.3	2.0	1.5	1.3	1.5	92.0	91.1	80.6	85.7	87.3
SynSegNet [21]	2.8	0.7	4.8	2.5	2.7	87.2	90.2	76.6	79.6	83.4
AdaOutput [29]	1.9	1.4	3.0	1.8	2.1	85.8	89.7	76.3	82.2	83.5
CycleGAN [10]	2.0	3.2	1.9	2.6	2.4	88.8	87.3	76.8	79.4	83.1
CyCADA [15]	1.5	1.7	1.3	1.6	1.5	88.7	89.3	78.1	80.2	84.1
SIFA-v1 [18]	2.3	0.9	1.4	2.4	1.7	88.5	90.0	79.7	81.3	84.9
SIFA-v2 [28]	1.5	0.6	1.5	2.4	1.5	90.0	89.1	80.2	82.3	85.4
AttENT	**0.99**	**1.03**	**1.26**	**1.12**	**1.10**	**91.05**	**81.38**	**80.51**	**89.75**	**85.67**
W/o Adapt	4.5	12.3	6.8	4.5	7.0	48.9	50.9	65.3	65.7	57.7

0 means no overlap. The calculation formula is:

$$Dice(A,B) = \frac{2 \times |A \cap B|}{|A| + |B|} = \frac{2 \times A \cdot B}{A^2 + B^2} \tag{9}$$

where $|A|$ represents the number of pixels in the A. So $|A \cap B|$ represents the overlap area contrasted pixel by pixel between ground true and predict result, which is usually approximated as the result of pixel-by-pixel accumulation after point multiplication between the prediction result and ground true.

ASSD determines the average difference between the surface of the predicted object and the ground truth. After the border voxels of segmentation and reference are determined, those voxels that have at least one neighbor from a predefined neighborhood that does not belong to the object are collected. For each collected voxel, the closest voxel in the other set is determined and the average of all these distances gives ASSD (0 for a perfect segmentation, max distance of prediction for the worst segmentation). The calculation formula is:

$$\begin{aligned} ASSD(A,B) = (&\sum_{a \in S(A)} \min_{b \in S(B)} \|a-b\| + \\ &\sum_{b \in S(B)} \min_{a \in S(A)} \|b-a\|)/\left(|S(A)| + |S(B)|\right) \end{aligned} \tag{10}$$

where $S(\ )$ indicates the set of pixels on the surface, $*$ means Euclidean Distance between two points, $|\ |$ represents the number of pixels in the set.

Higher Dice and lower ASSD mean the predicted result are more similar as the ground truth and fewer distances between the surface respectively.

3) Experimental implement details: For the deep learning environment, we built the Python with version 3.6.9 and PyTorch with version 1.3.0 on the Ubuntu system (18.04) with one NVIDIA 1080Ti GPU.

For image alignment in the pixel space, as shown in Fig.2, the CycleGAN [10] with inserted Attention Gates are designed as backbone. Generator G_{S2T} and G_{T2S} has the same architecture, which consist of encoder and decoder for feature extraction and reconstruction respectively. Discriminator D_S and D_T also have the same architecture, which stacked by four layers (Conv, InstanceNorm, LeakyReLU), and finally attach FCN to get the classification result. Above networks are trained by the Adam optimizer [32] with the initial learning rate 0.001. And the well-trained model is obtained after 10k iterations with batchsize 1.

For feature alignment in the entropy space, as shown in Fig.3, the segmentor S is built on DeepLab-v2 [33] and pretrained based on the ImageNet [34]. The discriminator D is stacked by four layers (Conv, LeakyReLU), and finally attach FCN to get the classification result. The entropy information is calculated based on Shannon Entropy [35]. Above networks are trained by the Adam optimizer [32] with the initial learning rate 0.0001. And the well-trained model is obtained after 20 epochs with batchsize 8.

4) ablation study: In order to demonstrate the effectiveness of key components in the proposed method, we performed ablation study for domain adaptation in the pixel space and entropy space respectively. The ablation study experiments were designed as follows.

Firstly, we conducted experiments without any domain adaptation and regarded these results as the baseline. Secondly, we respectively obtained the segmentation results of image alignment in pixel space and CycleGAN, feature alignment in the original feature space and entropy space. And then we compared the above results of four conditions in order:

- do the image translation with classic CycleGAN (analyzed in section II).
- do the image alignment with attention-aware CycleGAN (analyzed in section III-B).
- do the feature alignment in the feature space (analyzed in section II).
- do the feature alignment in the entropy space (analyzed in section III-C).

In the end, we merged image alignment in the pixel space and feature alignment in the entropy space to get AttENT for final validation. Meanwhile, we also conducted significance tests on the experimental results to evaluate the effectiveness of key components.

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

Firstly, in order to quantitatively measure the domain gap in the abdominal organs segmentation, we measured the domain shift by conducted experiments to get the performance gap before comparison. The bottom bound was to directly transfer the well-trained model from source domain to target domain. The top bound was to use annotated target images to carry out supervised learning of the target model. The numerical difference in performance values between the bottom bound and top bound is caused by the domain shift. After that, in order to fairly evaluate the difference, we selected six state-of-the-art UDA methods [10], [15], [18], [20], [21], [28] and followed the same settings as [28]. The quantitative results of domain-adaptive segmentation of abdominal organs are shown in the Table I, where the results of above SOTA methods all came from the paper [28].

More specifically, we can only get the average Dice of 58.2% when adaptating the well-trained model from source domain (MRI) to target domain (CT), and get the average Dice of 57.7% when transferring from source domain (CT) to target domain (MRI). The huge performance gap between top bound and bottom bound also indicated the severe domain shift between MRI and CT images, which increased the difficulty of domain-adaptive medical image segmentation. However, we made remarkable improvement by achieving the average Dice of 85.29% over the four organs with the average ASSD being reduced to 1.16. And for CT ⇒ MRI, our AttENT improved the average Dice to 85.67% and reduced the average ASSD to 1.10. In general, our proposed AttENT surpassed most existing SOTA UDA methods in both cross-domain segmentation tasks in terms of both Dice and ASSD metrics. For example, compared with SIFA [28], the proposed AttENT reduced the average ASSD by 40 percentage points for CT ⇒ MRI and increased the average Dice from 83.7% to 85.67% for MRI ⇒ CT. Besides, all above results were very close or even exceed the top bound of supervised training and also demonstrated the effectiveness of our method.

The qualitative segmentation results in Fig.4 also showed that it is difficult to obtain correct prediction for any abdominal structure without adaptation. Instead, our method can successfully locate the four organs and generate semantically

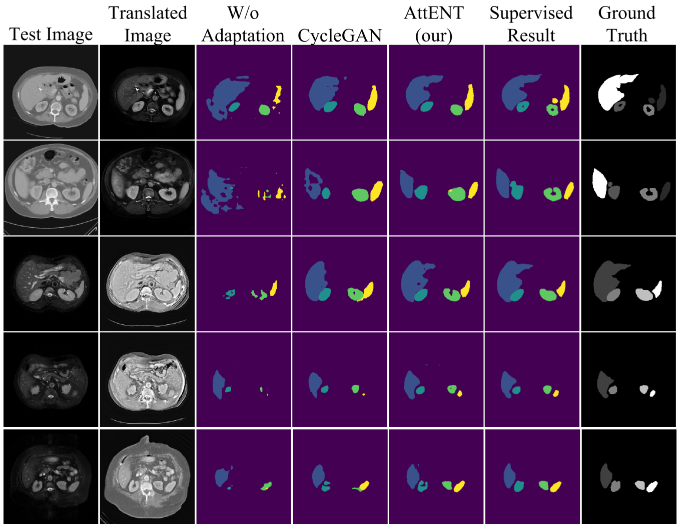


Fig. 4. **Visual comparison of segmentation results produced by different methods for abdominal organs segmentation.** CT images are shown in the top two rows and MRI images are shown in the bottom three rows. From left to right are the raw test images (1st column),translated images (2nd column) "W/o Adaptation" lower bound (3rd column), results of CycleGAN (4th column), results of our AttENT (5th column), results of supervised training (6th column), and ground truth (last column). The liver, right kidney, left kidney, and spleen are indicated in four different colors. Each row corresponds to one sample.

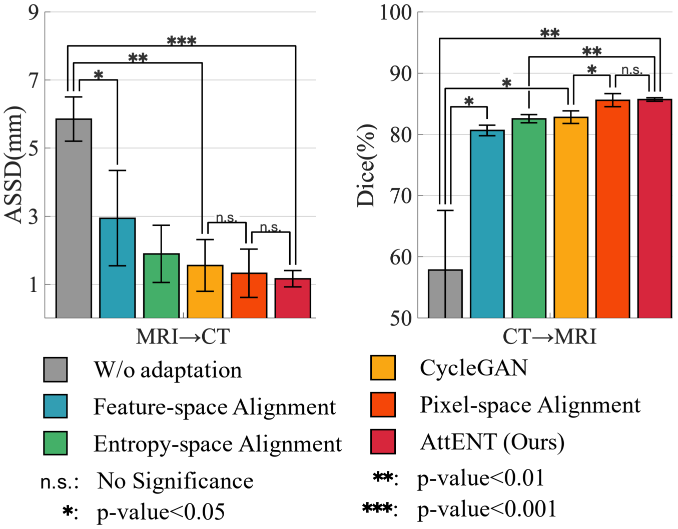


Fig. 5. **Ablation study of key components in our method.** The the mean±std of the experimental results are displayed as a histogram, and the results of significance tests are reported by the p-value.

TABLE II

The results of ablation experiments and the results of the significance tests. The values in the first row and the first column of each table represent the performance evaluation results of the method in the domain adaptation tasks, expressed in ASSD or Dice. The significance test result p-value is displayed in different rows and columns in the table, representing the significant difference level of the method performance indicated by the row and column. (\ means the insignificant comparison.)

MRI ⇒ CT						
Methods ASSD (Mean±Std.)	W/o Adaptation (5.85±0.65)	CycleGAN (1.55±0.76)	Pixel Space Alignment (1.32±0.71)	Feature Space Alignment (2.94±1.40)	Entropy Space Alignment (1.89±0.84)	AttENT (1.16±0.24)
W/o Adaptation (5.85±0.65)	\	0.0018	\	0.0281	\	0.0003
CycleGAN (1.55±0.76)	0.0018	\	0.3827	\	\	\
Pixel Space Alignment (1.32±0.71)	\	0.3827	\	\	\	0.3895
Feature Space Alignment (2.94±1.40)	0.0281	\	\	\	0.2063	\
Entropy Space Alignment (1.89±0.84)	\	\	\	0.2063	\	0.1521
AttENT (1.16±0.24)	0.0003	\	0.3895	\	0.1521	\

CT ⇒ MRI						
Methods Dice (Mean±Std.)	W/o adaptation (57.73±9.72)	CycleGAN (83.05±1.03)	Pixel Space Alignment (85.57±1.07)	Feature Space Alignment (80.62±0.86)	Entropy Space Alignment (82.54±0.67)	AttENT (85.67±0.29)
W/o Adaptation (57.73±9.72)	\	0.0113	\	0.0149	\	0.0078
CycleGAN (83.05±1.03)	0.0113	\	0.0283	\	\	\
Pixel Space Alignment (85.57±1.07)	\	0.0283	\	\	\	0.4524
Feature Space Alignment (80.62±0.86)	0.0149	\	\	\	0.0337	\
Entropy Space Alignment (82.54±0.67)	\	\	\	0.0337	\	0.0019
AttENT (85.67±0.29)	0.0078	\	0.4524	\	0.0019	\

meaningful mask. Both the quantitative results (Table I) and qualitative results (Fig.4) validate the effectiveness of our method on addressing the severe domain shift.

C. Effectiveness of Key Components

Table. II reported the performance discrepancy between different components of AttENT on multi-modalities abdominal organs segmentation. In addition to quantitative representation, we also conducted significance tests for some key components pair. The p-value of above comparison also reported in the Table. II. In order to obtain a more intuitive and clear demonstration, we showed the above ablation study results in the form of a histogram in Fig. 5.

We found that the performance of non-adaptive model got improved after introducing domain adaptation components. Besides, after synergizing pixel space alignment and entropy space alignment, our proposed AttENT achieved the best performance. Compared with the baseline (W/o Adaptation), AttENT significantly outperformed in the cross-modality medical image segmentation ($p\text{-value}<0.01$). More precisely, compared with the results of the image translation method with CycleGAN, ASSD reduced by 0.23±0.05 after adding the attention mechanism into pixel space when translating MRI to the CT, Dice showed a 2.52±0.04 percentage point growth when translating CT to the MRI. Similarly, compared with the results of only alignment in the feature space, the adopted entropy minimization principle reduced the ASSD by 1.05±0.56 for MRI ⇒ CT and improved the Dice by 1.92±0.19 percentage point for CT ⇒ MRI. The quantitative comparisons of performance verified the effectiveness of the proposed key components in Section III-B and III-C. Meanwhile, the significance tests verified the importance of synergizing pixel space alignment and entvropy space alignment to improve the performance of cross-modality medical image segmentation.

V. Conclusion

In this paper, a novel unsupervised domain-adaptive framework is proposed to recover performance degradation from the domain shift in cross-modality medical image segmentation. We synergize image alignment in pixel space and the feature alignment in entropy space for domain generalization. On the one hand, our framework is able to improve the image alignment by introducing the attention mechanism into CycleGAN in the pixel space. On the other hand, the principle of entropy-minimization is utilized to align features in the entropy space.

Finally, the proposed methodology learns the domain-adaptive representation and can effectively reduce the performance loss during cross-domain adaptation without any target annotations. Experiments demonstrate that in the domain-adaptive segmentation of cross-modality medical images on public datasets, our work reaches state-of-the-art.

Acknowledgment

This research was funded by the National Key Research and Development Program of China (No. 2018YFB0204301).

References

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2016, pp. 770–778.

[2] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2015, pp. 1–9.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems (NIPS)*, 2012, pp. 1097–1105.

[4] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Transactions on Medical Imaging (TMI)*, vol. 39, no. 6, pp. 1856–1867, 2020.

[5] C. Li, W. Chen, and Y. Tan, "Point-sampling method based on 3d u-net architecture to reduce the influence of false positive and solve boundary blur problem in 3d ct image segmentation," *Applied Sciences*, vol. 10, no. 19, 2020.

[6] P. Tang, C. Wang, X. Wang, W. Liu, W. Zeng, and J. Wang, "Object detection in videos by high quality object linking," *IEEE transactions on pattern analysis and machine intelligence (TPAMI)*, 2019.

[7] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2014, pp. 580–587.

[8] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[9] C. Li, W. Chen, and Y. Tan, "Render u-net: A unique perspective on render to explore accurate medical image segmentation," *Applied Sciences*, vol. 10, no. 18, 2020.

[10] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[11] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu, "Transfer feature learning with joint distribution adaptation," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2013.

[12] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng, "Self-taught learning: transfer learning from unlabeled data," in *Proceedings of the 24th International Conference on Machine Learning (ICML)*, 2007, pp. 759–766.

[13] N. Inoue, R. Furuta, T. Yamasaki, and K. Aizawa, "Cross-domain weakly-supervised object detection through progressive domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 5001–5009.

[14] E. Kodirov, T. Xiang, Z. Fu, and S. Gong, "Unsupervised domain adaptation for zero-shot learning," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

[15] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. Efros, and T. Darrell, "CyCADA: Cycle-consistent adversarial domain adaptation," in *Proceedings of the 35th International Conference on Machine Learning (ICML)*, vol. 80, 10–15 Jul 2018, pp. 1989–1998.

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

[17] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, vol. 37, 07–09 Jul 2015, pp. 1180–1189.

[18] C. Chen, Q. Dou, H. Chen, J. Qin, and P.-A. Heng, "Synergistic image and feature adaptation: Towards cross-modality domain adaptation for medical image segmentation," *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, vol. 33, pp. 865–872, 2019.

[19] Y. Tsai, W. Hung, S. Schulter, K. Sohn, M. Yang, and M. Chandraker, "Learning to adapt structured output space for semantic segmentation," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7472–7481.

[20] Q. Dou, C. Ouyang, C. Chen, H. Chen, and P.-A. Heng, "Unsupervised cross-modality domain adaptation of convnets for biomedical image segmentations with adversarial loss," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI)*, 2018, pp. 691–697.

[21] J. Jiang, Y.-C. Hu, N. Tyagi, P. Zhang, A. Rimner, G. S. Mageras, J. O. Deasy, and H. Veeraraghavan, "Tumor-aware, adversarial domain adaptation from ct to mri for lung cancer segmentation," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pp. 777–785.

[22] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems (NIPS)*, vol. 27, 2014, pp. 2672–2680.

[23] B. Chidlovskii, S. Clinchant, and G. Csurka, "Domain adaptation in the absence of source domain data," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2016, p. 451–460.

[24] R. Li, Q. Jiao, W. Cao, H.-S. Wong, and S. Wu, "Model adaptation: Unsupervised domain adaptation without source data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[25] O. Oktay, J. Schlemper, L. Le Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y Hammerla, B. Kainz *et al.*, "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

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

[27] T. Vu, H. Jain, M. Bucher, M. Cord, and P. Pérez, "Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 2512–2521.

[28] C. Chen, Q. Dou, H. Chen, J. Qin, and P. A. Heng, "Unsupervised bidirectional cross-modality adaptation via deeply synergistic image and feature alignment for medical image segmentation," *IEEE Transactions on Medical Imaging (TMI)*, vol. 39, no. 7, pp. 2494–2505, 2020.

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

[30] A. E. Kavur, N. S. Gezer, M. Barış, S. Aslan, and P.-H. Conze, "Chaos challenge - combined (ct-mr) healthy abdominal organ segmentation," *Medical Image Analysis*, vol. 69, p. 101950, 2021.

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

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

[33] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence (TPAMI)*, vol. 40, no. 4, pp. 834–848, 2017.

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

[35] J. Lin, "Divergence measures based on the shannon entropy," *IEEE Transactions on Information Theory (TIT)*, vol. 37, no. 1, pp. 145–151, 1991.