Unsupervised Domain Adaptation for Medical Image Segmentation by Selective Entropy Constraints and Adaptive Semantic Alignment

Wei Feng[1,2,3], **Lie Ju**[1,2,3], **Lin Wang**[1,2,3], **Kaimin Song**[4], **Xin Zhao**[4], **Zongyuan Ge**[1,2,3],*

[1]Monash eResearch Center, Monash University
[2]Monash Medical AI Group, Monash University
[3]Airdoc Monash Research Centre, Monash University
[4]Airdoc LLC
wei.feng@monash.edu, julie334600@gmail.com, wanglin.mailbox@gmail.com, kims75699@gmail.com, zhaoxin@airdoc.com, zongyuan.ge@monash.edu

Abstract

Generalizing a deep learning model to new domains is crucial for computer-aided medical diagnosis systems. Most existing unsupervised domain adaptation methods have made significant progress in reducing the domain distribution gap through adversarial training. However, these methods may still produce overconfident but erroneous results on unseen target images. This paper proposes a new unsupervised domain adaptation framework for cross-modality medical image segmentation. Specifically, We first introduce two data augmentation approaches to generate two sets of semantics-preserving augmented images. Based on the model's predictive consistency on these two sets of augmented images, we identify reliable and unreliable pixels. We then perform a selective entropy constraints: we minimize the entropy of reliable pixels to increase their confidence while maximizing the entropy of unreliable pixels to reduce their confidence. Based on the identified reliable and unreliable pixels, we further propose an adaptive semantic alignment module which performs class-level distribution adaptation by minimizing the distance between same class prototypes between domains, where unreliable pixels are removed to derive more accurate prototypes. We have conducted extensive experiments on the cross-modality cardiac structure segmentation task. The experimental results show that the proposed method significantly outperforms the state-of-the-art comparison algorithms. Our code and data are available at https://github.com/fengweie/SE_ASA.

Introduction

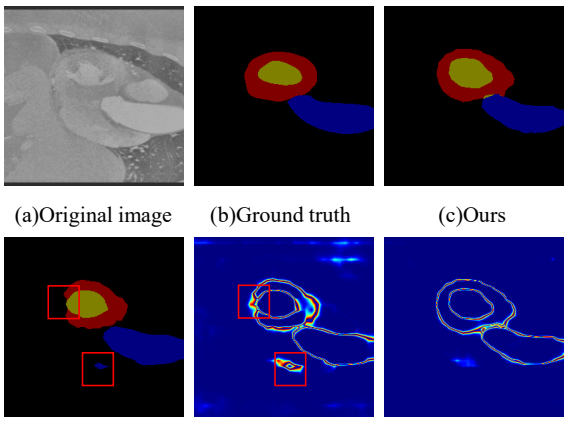
Deep learning techniques have achieved breakthroughs in various fields in recent years, including computer vision (Oquab et al. 2014; Li et al. 2021), natural language processing (Collobert and Weston 2008; Yin et al. 2017), speech recognition (Graves, Mohamed, and Hinton 2013; Abdel-Hamid et al. 2012), etc. However, when applied directly to an unseen dataset with a different distribution, a well-trained deep learning model often suffers from performance degradation due to distribution shifts (Djolonga et al. 2021). This phenomenon is even more common in the field of medical image analysis, where clinical data may have significantly different appearances as they are often acquired from different devices, different hospitals and different protocols. A naive solution would be to let the doctor annotate the the data of the new domain from scratch. However, in a clinical scenario, even for experienced physicians, annotating large amounts of medical image data is very time-consuming and labour-intensive.

Recently, unsupervised domain adaptation(UDA) methods have received increasing attention, with the aim of improving the generalization performance of a model on the target domain without re-annotating the target domain data, using only the annotated source domain data. Most of the current mainstream UDA methods use adversarial training to perform distribution adaptation in the image or feature space (Chen et al. 2020a; Tsai et al. 2018). Image-level UDA methods typically use image translation techniques (Zhu et al. 2017) to reduce the difference in appearance of images from the source and target domains. Feature-level UDA methods learn domain invariant feature representations by adversarial training (Luo et al. 2019; Feng et al. 2022). Considering that the spatial structure of the predicted outputs in the source and target domains should be similar, some methods perform distribution adaptation in the output space (Tsai et al. 2018).

However, due to the properties of cross-entropy loss, which forces the network output to match the one-hot ground truth label, the neural network may be mis-calibrated and would output overconfident predictions (Guo et al. 2017; Zou et al. 2019). This phenomenon is exacerbated in the presence of domain shifts, as the model only receives supervision signals from the source domain during training (Wang et al. 2020). For example, as shown in Fig. 1, the state-of-the-art algorithm SIFA V2 (Chen et al. 2020a) on the cross-modality cardiac segmentation dataset produced overconfident (low entropy) segmentation results on the target image, which required explicit correction. Most adversarial training-based UDA methods do not address this problem properly.

To address the above problem, in this paper we propose a new unsupervised domain adaptation algorithm based on **selective entropy constraints** and **adaptive semantic**

*Corresponding Author: zongyuan.ge@monash.edu.



(a)Original image (b)Ground truth (c)Ours

(d)SIFA v2 (e)SIFA v2 entropy map (f)Ours entropy map

Figure 1: Illustration of the source model producing incorrect but overconfident predictions on the target data (marked with red bounding boxes). (a) Original image. (b) Ground truth. (c) Predictions of our method. (d) Predictions of SOTA UDA method SIFA V2 on this cardiac structure segmentation dataset. (e) Entropy map of SIFA V2. (f) Entropy map of our method.

alignment. First, we develop two data augmentation methods, namely random transformation-based data augmentation and Fourier-based data augmentation to generate two sets of augmented images. We then identify the reliable and unreliable pixels based on the model's prediction consistency on the two augmented image sets. We then propose a selective entropy constraints strategy to remold the model predictions. Specifically, we consider those pixels with consistent predictions to be reliable and therefore minimize their prediction entropy to increase the confidence score; meanwhile, those pixels with inconsistent predictions may be unreliable and therefore we maximize their prediction entropy to decrease their confidence score. In addition, to further narrow category-level feature distribution differences between the source and target domains, we propose an adaptive semantic alignment module that minimizes the distance between the prototypes of the same categories in the source and target domains, where those unreliable pixels are removed to avoid domain confusion. The main contributions of this paper are summarized as follows:

- We propose a selective entropy constraints strategy which remolds model confidence for reliable and unreliable pixels identified based on predictive consistency on two augmented image sets.
- An adaptive semantic alignment module to reduce category-level distribution differences between domains, where unreliable pixels are removed to avoid domain confusion.
- We validate the performance of the proposed algorithm on the challenging domain adaptation task of cross-modality cardiac structure segmentation. The performance of our approach outperforms the state-of-the-art comparison algorithms by a large margin.

Related Work

Unsupervised Domain Adaptation: Addressing the performance degradation of deep learning models due to domain shifts by unsupervised domain adaptation has been a very active research topic in recent years. Distribution matching based on adversarial training is the dominant UDA approach. For example, Hoffman et al. (Hoffman et al. 2018) used generative adversarial networks to convert source images to target image styles and then trained models on these target-like images. Russo et al. (Russo et al. 2018) transformed target images to source image styles and then tested them using the source models. Tsai et al. (Tsai et al. 2018) performed adversarial training in the output space considering that the source and target predictions have similar spatial structure distribution. Chen et al. (Chen et al. 2019) combined image-level and feature-level adaptation into a unified framework and used it for cross-modality cardiac structure segmentation.

However, due to the lack of explicit constraints, models adapted after adversarial training may still produce overconfident, erroneous predictions in some regions of the target domain images. In contrast to these methods, our approach is able to selectively impose constraints on the predictions of reliable and unreliable pixels, and thus effectively alleviating the overconfidence problem. In addition, based on these two sets of pixels, we also propose an adaptive semantic alignment module to achieve category-level distribution adaptation.

Confidence Regularization: Confidence regularization is a widely adopted technique in supervised learning to alleviate overfitting effects. Commonly used confidence regularization techniques include label smoothing (Liu et al. 2019), temperature scaling (Guo et al. 2017), and network output regularization (Pereyra et al. 2017), etc. In addition, a few studies have also explored confidence regularization in UDA scenarios. For example, Liu et al. (Liu et al. 2021a) propose a pseudo-label-independent energy-based model to constrain the training of target domain samples. Zou et al. (Zou et al. 2019) propose a confidence regularization-based self-training method and demonstrated that both label regularization and model regularization can alleviate the overconfidence problem to some extent. In contrast to these approaches, we identify reliable and unreliable pixels based on prediction consistency and perform adaptive confidence adjustment to improve model performance.

Prediction Consistency: The prediction consistency regularization technique has been shown to be beneficial in many scenarios such as unsupervised domain adaptation (Ma et al. 2021), semi-supervised learning (Berthelot et al. 2019; Sohn et al. 2020) and self-supervised learning (Chen et al. 2020b). In addition, Bahat et al. (Bahat, Irani, and Shakhnarovich 2019) found that the reliability of a sample can be identified by the consistency of the model's predictions under data augmentation. In this paper, we use the pre-

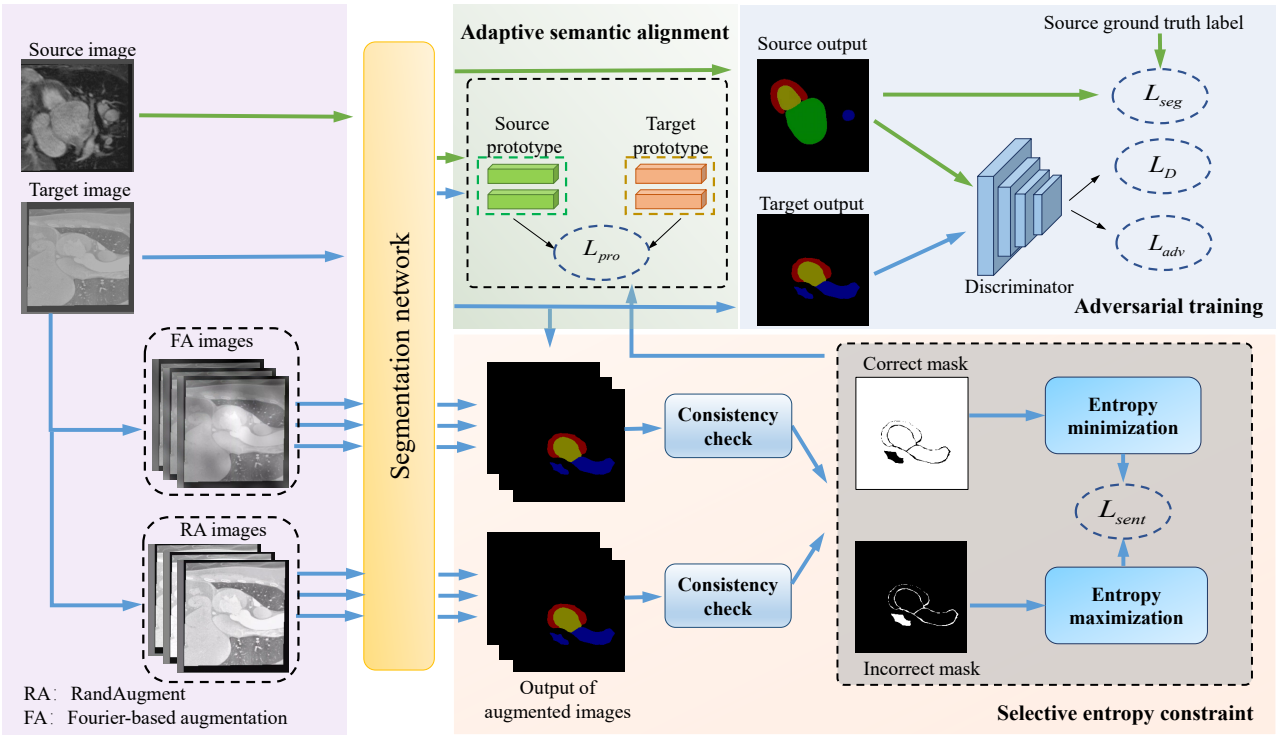


Figure 2: An overview of the proposed unsupervised domain adaptation framework. We identify reliable and unreliable pixels based on prediction consistency, and then apply selective entropy constraints to both. At the same time, we perform global distribution adaptation by adversarial training. Adaptive semantic alignment is performed by minimizing the distance between same-category prototypes between domains. We train the model in an end-to-end fashion.

diction consistency of the model on two augmented image sets to identify the reliability of pixels.

Methodology

Fig. 2 illustrates the overall framework of the proposed unsupervised domain adaptive algorithm. We use two data augmentation strategies to obtain augmented images, and then perform selective entropy constraints based on the consistency of the model's predictions on the augmented images. We also combine adaptive semantic alignment and adversarial training to perform feature distribution adaptation.

Problem Definition

In the traditional UDA setup, we are given a labelled source domain dataset denoted as $\mathcal{D}^s = \{(x_i^s, y_i^s)\}_{i=1}^{N_s}$, where x_i^s is the ith source domain image, $y_i^s \in \{1, 2, \ldots, C\}$ is its corresponding pixel-level label. In addition, we also have access to an unlabelled target domain dataset $\mathcal{D}^t = \{x_i^t\}_{i=1}^{N_t}$. To encourage the output of the source and target domains to have a similar spatial structure, previous approach introduces the adversarial training strategy, which can be formulated as a min-max game (Tsai et al. 2018). Specifically, we construct a generator and a discriminator, with the discriminator being used to distinguish between source and target domain features, while the generator is trained to fool the discriminator. The optimization objective for the discriminator is:

$$L_D = \frac{1}{N_s}\sum_{i=1}^{N_s} L_{bce}\left(p_i^s, 1\right) + \frac{1}{N_t}\sum_{i=1}^{N_t} L_{bce}\left(p_i^t, 0\right), \quad (1)$$

where L_{bce} is the binary cross-entropy loss and p_i^s and p_i^t are the predicted probability maps for the source and target domains. The adversarial loss on the generator is defined as:

$$L_{adv} = \frac{1}{N_t}\sum_{i=1}^{N_t} L_{bce}\left(p_i^t, 1\right) \quad (2)$$

However, deep learning models are often poorly calibrated and may output overconfident predictions especially in the presence of distribution shifts. Although the domain distribution discrepancy can be reduced by adversarial training, the overconfidence problem is not addressed, see Fig. 1.

Selective Entropy Constraints

To address the above problem, we propose a selective entropy constraints strategy to remold the prediction confidence of the model. Specifically, we first propose two data augmentation methods to obtain two sets of augmented target images.

Random Transformation-Based Data Augmentation Given a target image x_i^t, we apply a series of random image

transformations, such as color jittering and Gaussian blur (Chen et al. 2020b), to generate N augmented images based on random transformations:

$$x_{ik}^{RA} = RandAugment_k(x_i^t), k = 1, \ldots, N \quad (3)$$

Fourier-Based Data Augmentation To further increase the diversity of image transformations, we introduce a Fourier-based data augmentation strategy. The Fourier transform has been found to be an effective data augmentation method for domain generalization (Xu et al. 2021), domain adaptation tasks (Yang and Soatto 2020). Unlike previous methods, we use predictive consistency on Fourier transform-based augmented images to assess the reliability of the pixels. Specifically, for each target image x_i^t, we randomly select N source-domain images and then transform these N source images into the frequency domain using the Fourier transformation to obtain the amplitude component $\{\mathcal{A}_k^s\}_{k=1}^N$ and the phase component $\{\mathcal{P}_k^s\}_{k=1}^N$. Similarly, we obtain the amplitude component $\mathcal{A}_i^t$ and the phase component $\mathcal{P}_i^t$ of the target image. Since the amplitude component contains the low-level statistics of the image and the phase component contains the high-level semantic information of the image (Xu et al. 2021), we linearly interpolate the amplitude components of the target and source images to obtain the new amplitude components:

$$\mathcal{A}_{ik}^{FA} = (1-\mu)\mathcal{A}_i^t * (1-\mathcal{V}) + \mu\mathcal{A}_k^s * \mathcal{V}, k = 1, \ldots, N \quad (4)$$

where μ controls the degree of augmentation of the amplitude components and $\mathcal{V}$ is a binary mask used to control the region of the amplitude spectrum to be swapped, which we set to be the central region containing the low-level information (Yao, Hu, and Li 2022). We then combine the augmented amplitude and phase components and use the inverse Fourier transformation $\mathcal{F}^{-1}$ to obtain the augmented images:

$$x_{ik}^{FA} = \mathcal{F}^{-1}\left(\mathcal{A}_{ik}^{FA}, \mathcal{P}_i^t\right), k = 1, \ldots, N \quad (5)$$

Based on these two sets of augmented images, we measure the consistency of the model's predictions over the original image and the augmented images in each image set separately. We then select pixels using a simple but effective majority voting mechanism: a pixel is considered reliable if a majority of the model's predictions on the augmented version of the pixel reach a consensus with the model's predictions on the original pixel. Conversely, the pixel is considered unreliable. We can then obtain two consistency masks C_i^{RA}, C_i^{FA} and two inconsistency masks I_i^{RA}, I_i^{FA}. When $C_{ij}^{RA} \wedge C_{ij}^{FA}$, we finally judge the jth pixel to be reliable; when $I_{ij}^{RA} \vee I_{ij}^{FA}$, the jth pixel is finally judged to be unreliable.

After obtaining the reliable and unreliable pixels, we perform the selective entropy constraints. Specifically, we minimize the prediction entropy of those reliable pixels to increase their confidence, and maximize the prediction entropy of those unreliable pixels to decrease their confidence, avoiding the model from producing overconfident predictions, which can be formulated as:

$$L_{sent} = \begin{cases} +L_{cem}\left(x^t\right), & \text{if} \quad C_{ij}^{RA} \wedge C_{ij}^{FA} \\ -L_{cem}\left(x^t\right), & \text{if} \quad I_{ij}^{RA} \vee I_{ij}^{FA} \end{cases} \quad (6)$$

$$L_{cem} = -\sum_{i=1}^{N_t}\sum_{j=1}^{H\times W}\sum_{c=1}^{C} p_{ijc}^t \log\left(p_{ijc}^t\right). \quad (7)$$

Adaptive Semantic Alignment

With the selective entropy constraints strategy, we can regularize the confidence of the model during the training process. However, category level distribution mismatches would still result in the adapted model not generalizing well on the target domain. Ideally, after adaptation, features of pixels of the same category from different domains should be mapped nearby. Here, we propose an adaptive semantic alignment module that minimizes the distance between the same category prototypes in the labelled source domain and the pseudo-labelled target domain. We first calculate the prototypes for each category in the source and target domains:

$$\begin{aligned} z_c^s &= \frac{1}{|\Phi_c^s|}\sum_v \mathbb{I}_{[y_v^s=0]} e_v^s, \\ z_c^t &= \frac{1}{|\Phi_c^t|}\sum_v \mathbb{I}_{[\widehat{y}_v^t=0]} e_v^t \end{aligned} \quad (8)$$

where e denotes the output feature map of the penultimate layer of the model, Φ_c represent the pixels in e that belong to the cth class. $|\cdot|$ is the number of pixels in the pixel set. $\mathbb{I}$ is the indicator function. $\widehat{y}^t = argmax(p^t)$ are the pseudo labels of the target domain images. Note that since our model is actually optimized using the min-batch SGD algorithm, the class information in each min-batch may not be sufficient. Therefore, inspired by (Xie et al. 2018), we maintain a global prototype instead of the prototype in the current batch. Specifically, we first initialize the prototype with a forward computation, and then update the prototype in each iteration based on the data from the current batch as follows:

$$\begin{aligned} z_c^s &\leftarrow \alpha z_c^s + \frac{1-\alpha}{|\Phi_c^s|}\sum_v \mathbb{I}_{[y_v^s=1]} e_v^s, \\ z_c^t &\leftarrow \alpha z_c^t + \frac{1-\alpha}{|\Phi_c^t|}\sum_v \mathbb{I}_{[\widehat{y}_v^t=1]} e_v^t, \end{aligned} \quad (9)$$

where α is the coefficient used to update the prototype.

After calculating the prototypes for the source and target domains, we minimize the distance between prototypes of the same categories between domains, which can be formulated as:

$$L_{pro} = \sum_{c=1}^{C}\left\|z_c^s - z_c^t\right\|_2, \quad (10)$$

with the semantic alignment loss, we are able to maintain semantic consistency in the feature space.

However, due to differences in domain distribution, pseudo labels generated using model trained with labelled

Cardiac CT → Cardiac MRI										
Method	Dice					ASD				
	AA	LAC	LVC	MYO	Average	AA	LAC	LVC	MYO	Average
Supervised training	81.6	86.3	92.3	80.0	85.1	3.4	2.1	1.7	1.6	2.2
W/o adaptation	18.5	7.3	53.5	2.1	20.4	7.1	25.8	8.7	29.9	17.9
PnP-AdaNet (Dou et al. 2019)	43.7	47.0	77.7	48.6	54.3	11.4	14.5	4.5	5.3	8.9
AdaOutput (Tsai et al. 2018)	52.3	71.7	79.5	49.2	63.2	9.0	**3.5**	5.1	5.4	5.8
CycleGAN (Zhu et al. 2017)	64.3	30.7	65.0	43.0	50.7	5.8	9.8	6.0	5.0	6.6
CyCADA (Hoffman et al. 2018)	60.5	44.0	77.6	47.9	57.5	7.7	13.9	4.8	5.2	7.9
SIFA V1 (Chen et al. 2019)	67.0	60.7	75.1	45.8	62.1	6.2	9.8	4.4	4.4	6.2
SIFA V2 (Chen et al. 2020a)	65.3	62.3	78.9	47.3	63.4	7.3	7.4	**3.8**	4.4	5.7
EBM (Liu et al. 2021a)	65.9	64.2	76.9	49.1	64.1	6.9	7.5	5.6	3.8	6.0
CRST(Zou et al. 2019)	65.1	66.9	77.2	50.0	64.8	6.4	6.3	5.5	4.0	5.6
Ours	**68.3**	**74.6**	**81.0**	**55.9**	**69.9**	**4.9**	3.6	5.4	**3.2**	**4.3**

Cardiac MRI → Cardiac CT										
Method	Dice					ASD				
	AA	LAC	LVC	MYO	Average	AA	LAC	LVC	MYO	Average
Supervised training	89.3	91.4	92.8	88.0	90.4	2.3	2.9	1.5	3.2	2.5
W/o adaptation	30.8	36.8	18.3	7.2	23.3	20.2	8.9	33.6	27.8	22.6
PnP-AdaNet (Dou et al. 2019)	74.0	68.9	61.9	50.8	63.9	12.8	6.3	17.4	14.7	12.8
AdaOutput (Tsai et al. 2018)	73.5	80.4	76.1	48.6	69.6	15.5	5.8	5.2	6.6	8.3
CycleGAN (Zhu et al. 2017)	73.8	75.7	52.3	28.7	57.6	11.5	13.6	9.2	8.8	10.8
CyCADA (Hoffman et al. 2018)	72.9	77.0	62.4	45.3	64.4	9.6	8.0	9.6	10.5	9.4
SIFA V1 (Chen et al. 2019)	81.1	76.4	75.7	58.7	73.0	10.6	7.4	6.7	7.8	8.1
SIFA V2 (Chen et al. 2020a)	81.3	79.5	73.8	61.6	74.1	**7.9**	6.2	5.5	8.5	7.0
EBM (Liu et al. 2021a)	78.9	80.7	75.7	60.5	74.0	8.6	6.6	4.7	8.2	7.1
CRST(Zou et al. 2019)	79.6	80.5	78.3	63.7	75.5	8.8	6.4	4.5	7.5	6.8
Ours	**83.8**	**85.2**	**82.9**	**71.7**	**80.9**	9.6	**4.2**	**3.9**	**3.9**	**5.4**

Table 1: Performance of different domain adaptation algorithms for cardiac structure segmentation task

source data would have some incorrect predictions and using these noisy pixels to calculate prototypes would affect semantic alignment (Liu et al. 2021b). It is therefore necessary to filter out these noisy and unreliable pixels, and thanks to the reliable and unreliable pixels identified earlier according to our consistency strategy, we can easily remove the unreliable pixels and use only the reliable ones to compute more accurate prototypes for adaptive semantic alignment.

In summary, the optimization objective of the segmentation network can be formulated as follows:

$$L_{total} = L_{seg} + \lambda_1 L_{adv} + \lambda_2 L_{pro} + \lambda_3 L_{sent} \qquad (11)$$

where $L_{seg} = \sum_{i=1}^{N_s} \left(L_{CE}\left(y_i^s, p_i^s\right) + L_{Dice}\left(y_i^s, p_i^s\right)\right)$ is the supervised loss on the labeled source images (Chen et al. 2020a), $\lambda_1, \lambda_2, \lambda_3$ are balance coefficients.

Experiments and Results

Dataset. To validate the effectiveness of the proposed algorithm, we conducted experiments on the widely used Multi-Modality Whole Heart Segmentation (MMWHS) Challenge 2017 dataset (Zhuang and Shen 2016). The dataset contains 20 unpaired CT volumes and 20 MRI volumes from different sites. We segment four cardiac structures: the ascending aorta (AA), the left atrium blood cavity (LAC), the left ventricle blood cavity (LVC) and the myocardium of the left ventricle (MYO). We perform the adaptation task in two directions, from CT to MRI and from MRI to CT. For a fair comparison, we use the pre-processed data released by SIFA V2 (Chen et al. 2020a), of which 80% is used for training and 20% for testing. To evaluate the segmentation performance of the models, we used two widely used evaluation metrics: Dice coefficient(Dice) and the average symmetric surface distance(ASD) (Chen et al. 2020a). Dice measures the degree of overlap between model predictions and ground truth labels, and ASD measures the model's segmentation performance at the surface. Higher Dice values and lower ASD values indicate better segmentation performance.

Implementation Details and Evaluation Metrics. We chose DeepLabV2 (Chen et al. 2017) initialized with pretrained parameters on ImageNet (Deng et al. 2009) as the segmentation model, with the discriminator following the PatchGAN (Isola et al. 2017) setup. We used the stochastic gradient descent optimizer to train the segmentation model with learning rate set to 2.5×10^{-4}, momentum of 0.9 and weight decay of 10^{-4}. We used the Adam optimizer to train the discriminator with learning rate set to 10^{-4}. Similar to (Chen et al. 2020a), we performed distribution adaptation on the outputs of multiple levels, i.e. $conv4$ and $conv5$. The batch size is 4 and the number of iterations is 50,000. For

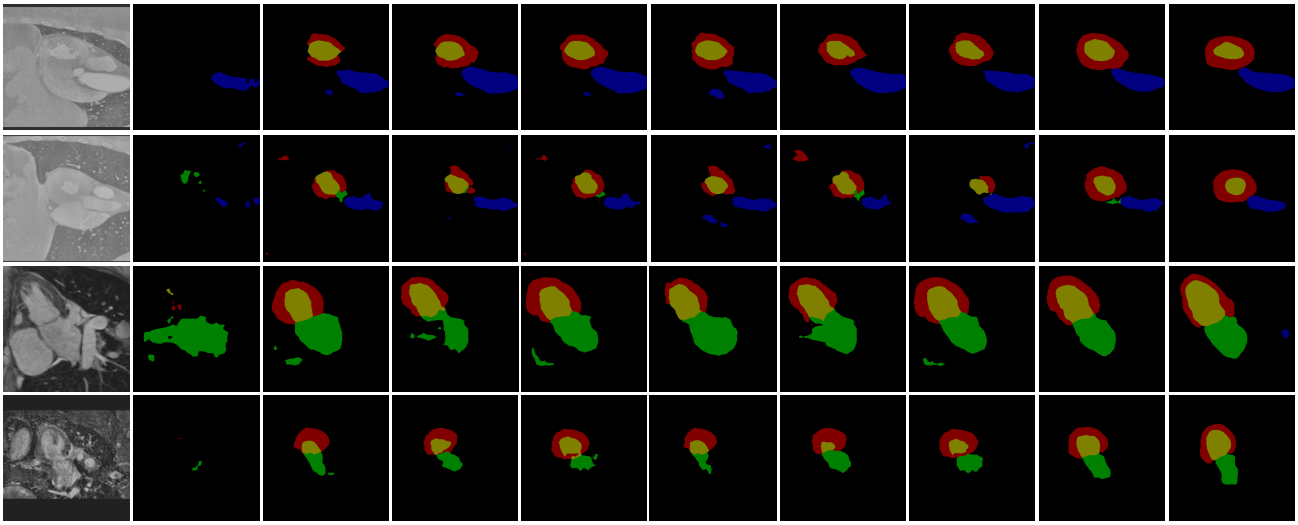


Figure 3: Visualization of segmentation results for different comparison algorithms. The top two rows show the segmentation results of the different algorithms on CT images and the bottom two rows show the segmentation results of the different algorithms on MRI images. The four cardiac structures AA, LAC, LVC and MYO are represented by blue, green, yellow and red respectively.

both augmentation methods, the number of augmentations N is set to 3.The prototype update coefficient α is set as 0.01, and $\lambda_1, \lambda_2, \lambda_3$ are set as 0.003, 0.1 and 1.0. The μ in Eq. 4 is set to 0.8. All models are implemented using PyTorch with 4 $\times$ 3090Ti GPUs.

Comparison with the State-of-the-Art Algorithms. To verify the superiority of the proposed algorithm, we compare it with a series of state-of-the-art comparison algorithms, including: (i) UDA methods based on image-level or feature-level adversarial training, i.e. PnP-AdaNet (Dou et al. 2019), AdaOutput (Tsai et al. 2018), CycleGAN (Zhu et al. 2017), CyCADA (Hoffman et al. 2018), SIFA V1 (Chen et al. 2019) and SIFA V2 (Chen et al. 2020a) (ii) Confidence regularized UDA methods, i.e. EBM (Liu et al. 2021a) and CRST (Zou et al. 2019). For SIFA V1 and SIFA V2, since they used the same dataset, we report the results from their papers directly. For the other comparison algorithms, we use their released code or re-implement them, noting that we use the same network structure to maintain a fair comparison. We also report the results without adaptation as the lower performance bound and the fully supervised results as the upper performance bound.

Tabel 1 lists the segmentation performance of the different algorithms on the two domain adaptive tasks CT $\rightarrow$ MRI and MRI $\rightarrow$ CT. It can be observed that the without adaptation model only obtains the average Dice of 20.4% and the average ASD of 17.9% on MRI images and the average Dice of 23.3% and the average ASD of 22.6% on CT images. This indicates that there are significant domain differences between CT and MRI images. It is worth noting that our method is significantly superior to other adversarial training-based UDA methods and confidence regularization

L_{adv}	L_{sent}	L_{pro}	Average Dice
✓			61.9
✓	✓		65.1
✓		✓	64.5
✓	✓	✓	**69.9**

Table 2: Ablation study of key components on the cardiac CT $\rightarrow$ MRI adaptation task.

EntMin	EntMax	Average Dice
all pixels	None	62.5
reliable pixels	None	64.7
reliable pixels	unreliable pixels	**69.9**

Table 3: Ablation study for the selective entropy constraints.

UDA methods. For example, compared with the best comparison algorithm, CRST, we obtained 5.1% average Dice value improvement and 1.3% average ASD value improvement on MRI images and 5.4% average Dice value improvement and 1.4% average ASD value improvement on CT images. Fig. 3 visualizes a comparison of the segmentation results of the different algorithms on CT and MRI images. It can be seen that W/o adaptation has difficulty in correctly segmenting the cardiac structures due to the domain shift problem. At the same time, our method is able to obtain segmentation results that are closer to the ground truth labels compared to other comparison algorithms.

Ablation Study. We first verify the effectiveness of the key components of the proposed algorithm. Our benchmark algorithm uses only adversarial training for distri-

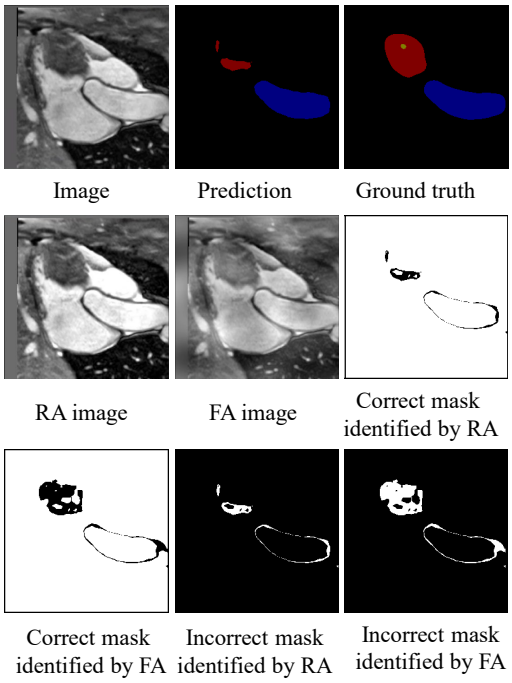


Figure 4: Illustration of pixel reliability assessment based on prediction consistency. RA,FA denote random transformation-based data augmentation and Fourier-based data augmentation, respectively.

Method	Average Dice
do not remove	66.5
remove	**69.9**

Table 4: Impact of removing unreliable pixels on adaptive semantic alignment.

bution adaptation. As shown in Tabel 2, the selective entropy constraints and adaptive semantic alignment result in a 3.2% and 2.6% Dice improvement, respectively, indicating that these two modules are useful for cross-modality medical image segmentation. Furthermore, our method achieves the best segmentation performance by combining these two components. We further investigate the effectiveness of different components of the selective entropy constraints. As shown in Tabel 3, it can be seen that the performance of the unconstrained condition entropy minimization method is 7.4% worse than our method! This demonstrates the importance of selective entropy constraints. Furthermore, we only perform entropy minimization for reliable pixels, but without entropy maximization for unreliable ones, and it can be seen that the performance is 5.2% lower than ours. This suggests the need to remold the prediction confidence for unreliable pixels. In addition, we also investigate the need to remove unreliable pixels in adaptive semantic alignment. As shown in Tabel 4, the performance of semantic alignment with the prototype obtained based on all pixels is 3.4% lower than our method, demonstrating the need to remove those unreliable pixels to avoid affecting semantic alignment.

Visualization of Reliable and Unreliable Pixels. To provide a more intuitive understanding of the consistency-based pixel reliability assessment strategy, we visualize the reliable and unreliable pixels identified based on the two data augmentation methods, and we also show example images under the two data augmentations. As shown in Fig. 4, there are some erroneous, overconfident regions in the model's predictions for the original images. Through the consistency of the model's predictions on the two sets of augmented images, we were able to identify reliable and unreliable pixels and then remold the confidence of the model by selective entropy constraints. Furthermore, it can be observed that the two data augmentation methods are complementary and can mutually reinforce each other to help identify these regions.

Conclusion

In this paper, we propose a new unsupervised domain adaptation framework for cross-modality medical image segmentation. It identifies reliable and unreliable pixels by the predictive consensus of the model on augmented sets of target images, and then remolds the model prediction confidence values using selective entropy constraints. The adaptive semantic alignment module performs class-level feature distribution alignment to reduce domain gap. we conducted extensive experiments on the cross-modality cardiac structure segmentation task. The experimental results show that the performance of the proposed algorithm outperforms other comparison algorithms by a large margin on all metrics. The approach is effective and can be generalized to other unsupervised medical image segmentation tasks.

References

Abdel-Hamid, O.; Mohamed, A.-r.; Jiang, H.; and Penn, G. 2012. Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition. In *2012 IEEE international conference on Acoustics, speech and signal processing (ICASSP)*, 4277–4280. IEEE.

Bahat, Y.; Irani, M.; and Shakhnarovich, G. 2019. Natural and Adversarial Error Detection using Invariance to Image Transformations. arXiv:1902.00236.

Berthelot, D.; Carlini, N.; Goodfellow, I.; Papernot, N.; Oliver, A.; and Raffel, C. A. 2019. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32.

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. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 865–872.

Chen, C.; Dou, Q.; Chen, H.; Qin, J.; and Heng, P. A. 2020a. 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.

Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4): 834–848.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020b. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PMLR.

Collobert, R.; and Weston, J. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, 160–167.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Djolonga, J.; Yung, J.; Tschannen, M.; Romijnders, R.; Beyer, L.; Kolesnikov, A.; Puigcerver, J.; Minderer, M.; D'Amour, A.; Moldovan, D.; et al. 2021. On robustness and transferability of convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16458–16468.

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.

Feng, W.; Wang, L.; Ju, L.; Zhao, X.; Wang, X.; Shi, X.; and Ge, Z. 2022. Unsupervised Domain Adaptive Fundus Image Segmentation with Category-Level Regularization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 497–506. Springer.

Graves, A.; Mohamed, A.-r.; and Hinton, G. 2013. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, 6645–6649. Ieee.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On calibration of modern neural networks. In *International conference on machine learning*, 1321–1330. PMLR.

Hoffman, J.; Tzeng, E.; Park, T.; Zhu, J.-Y.; Isola, P.; Saenko, K.; Efros, A.; and Darrell, T. 2018. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, 1989–1998. Pmlr.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1125–1134.

Li, Z.; Liu, F.; Yang, W.; Peng, S.; and Zhou, J. 2021. A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE transactions on neural networks and learning systems*.

Liu, X.; Hu, B.; Liu, X.; Lu, J.; You, J.; and Kong, L. 2021a. Energy-constrained self-training for unsupervised domain adaptation. In *2020 25th International Conference on Pattern Recognition (ICPR)*, 7515–7520. IEEE.

Liu, X.; Zou, Y.; Che, T.; Ding, P.; Jia, P.; You, J.; and Kumar, B. 2019. Conservative wasserstein training for pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8262–8272.

Liu, Y.; Deng, J.; Gao, X.; Li, W.; and Duan, L. 2021b. BAPA-Net: Boundary adaptation and prototype alignment for cross-domain semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8801–8811.

Luo, Y.; Zheng, L.; Guan, T.; Yu, J.; and Yang, Y. 2019. 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*, 2507–2516.

Ma, H.; Lin, X.; Wu, Z.; and Yu, Y. 2021. Coarse-to-fine domain adaptive semantic segmentation with photometric alignment and category-center regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4051–4060.

Oquab, M.; Bottou, L.; Laptev, I.; and Sivic, J. 2014. Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1717–1724.

Pereyra, G.; Tucker, G.; Chorowski, J.; Łukasz Kaiser; and Hinton, G. 2017. Regularizing Neural Networks by Penalizing Confident Output Distributions. arXiv:1701.06548.

Russo, P.; Carlucci, F. M.; Tommasi, T.; and Caputo, B. 2018. From source to target and back: symmetric bi-directional adaptive gan. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8099–8108.

Sohn, K.; Berthelot, D.; Carlini, N.; Zhang, Z.; Zhang, H.; Raffel, C. A.; Cubuk, E. D.; Kurakin, A.; and Li, C.-L. 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33: 596–608.

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. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7472–7481.

Wang, X.; Long, M.; Wang, J.; and Jordan, M. 2020. Transferable calibration with lower bias and variance in domain adaptation. *Advances in Neural Information Processing Systems*, 33: 19212–19223.

Xie, S.; Zheng, Z.; Chen, L.; and Chen, C. 2018. Learning semantic representations for unsupervised domain adaptation. In *International conference on machine learning*, 5423–5432. PMLR.

Xu, Q.; Zhang, R.; Zhang, Y.; Wang, Y.; and Tian, Q. 2021. A fourier-based framework for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14383–14392.

Yang, Y.; and Soatto, S. 2020. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition, 4085–4095.

Yao, H.; Hu, X.; and Li, X. 2022. Enhancing Pseudo Label Quality for Semi-Supervised Domain-Generalized Medical Image Segmentation. arXiv:2201.08657.

Yin, W.; Kann, K.; Yu, M.; and Schütze, H. 2017. Comparative Study of CNN and RNN for Natural Language Processing. arXiv:1702.01923.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2223–2232.

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.

Zou, Y.; Yu, Z.; Liu, X.; Kumar, B.; and Wang, J. 2019. Confidence regularized self-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5982–5991.