O2M-UDA: Unsupervised dynamic domain adaptation for one-to-multiple medical image segmentation

Z. Jiang, Y. He, S. Ye, P. Shao, X. Zhu, Y. Xu, Y. Chen, J.-L. Coatrieux, S. Li, G. Yang

O2M-UDA: Unsupervised Dynamic Domain Adaptation for One-to-Multiple Medical Image Segmentation

Ziyue Jiang[a], Yuting He[a], Shuai Ye[a], Pengfei Shao[b], Xiaomei Zhu[c], Yi Xu[c], Yang Chen[a], Jean-Louis Coatrieux[d], Shuo Li[e], Guanyu Yang[a,*]

[a]*LIST, Key Laboratory of Computer Network and Information Integration (Southeast University), Ministry of Education, Nanjing, China*
[b]*Dept. of Urology, the First Affiliated Hospital of Nanjing Medical University, Nanjing, China*
[c]*Dept. of Radiology, the First Affiliated Hospital of Nanjing Medical University, Nanjing, China*
[d]*Univ Rennes, Inserm, LTSI - UMR1099, Rennes, F-35000, France*
[e]*Dept. of Medical Biophysics, University of Western Ontario, London, ON, Canad*

Abstract

One-to-multiple medical image segmentation aims to directly test a segmentation model trained with the medical images of a one-domain site on those of a multiple-domain site, suffering from segmentation performance degradation on multiple domains. This process avoids additional annotations and helps improve the application value of the model. However, no successful one-to-multiple unsupervised domain adaptation (O2M-UDA) work has been reported in one-to-multiple medical image segmentation due to its inherent challenges: distribution differences among multiple target domains (among-target differences) caused by different scanning equipment and distribution differences between one source domain and multiple target domains (source-target differences). In this paper, we propose an O2M-UDA framework called dynamic domain adaptation (DyDA), for one-to-multiple medical image segmentation, which has two innovations: **1)** *dynamic credible sample strategy (DCSS)* dynamically extracts credible samples from the target site and iteratively updates their number, thus iteratively expanding the generalization boundary of the model and minimizing the among-target differences; **2)** *hybrid uncertainty learning (HUL)* reduces the voxel-level and domain-level uncertainty simultaneously, thus minimizing the source-target differences from the detail and entire perspective concurrently. Experiments on two one-to-multiple medical image segmentation tasks have been conducted to demonstrate the performance of the proposed DyDA. The proposed DyDA achieved competitive segmentation results and high adaptation with an average of 83.8% and 48.1% dice for the two tasks, respectively, which has improved by 21.7% and 9.2% compared with no adaptation, respectively. The code developed in this study code can be downloaded at https://github.com/ZoeyJiang/DyDA.

Keywords: One-to-multiple medical image segmentation, One-to-multiple domain adaptation, Dynamic credible sample strategy, Hybrid uncertainty learning

*Corresponding author
Email address: yang.list@seu.edu.cn (Guanyu Yang)

2010 MSC: 00-01, 99-00

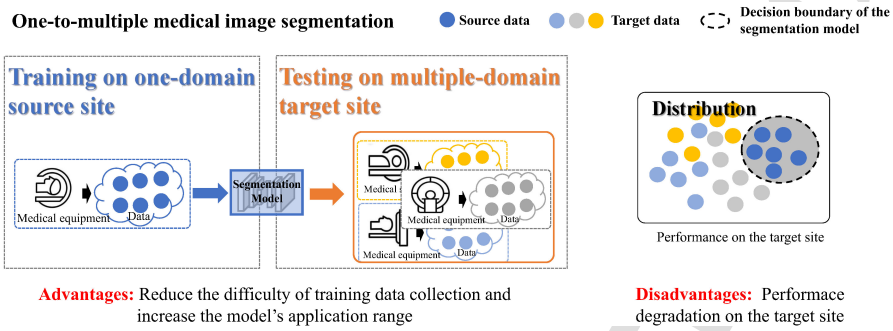

Figure 1: Illustration of what is one-to-multiple medical image segmentation. a) One-to-multiple medical image segmentation aims to directly test the segmentation model trained with the medical images of a one-domain source site on the medical images of a multiple-domain target site. b) One-to-multiple medical image segmentation reduces the difficulty of training data collection and improves the model's application range but suffers from segmentation performance degradation on the multiple-domain target site.

1. Introduction

One-to-multiple medical image segmentation is a critical medical image segmentation task of great clinical significance. As shown in Fig. 1, this task aims to test the segmentation model trained with the medical images of a one-domain site on the medical images of a multiple-domain site. Once successful, the segmentation model can be applied in medical images from different acquisition scenarios and equipment without additional annotations, thus improving the application value of the model and making medical images more valuable in the open environment [1]. **However**, as shown in Fig. 1, the segmentation model will suffer from segmentation performance degradation when tested on the target site directly due to cross-domain distribution differences [2, 3, 4] caused by different scanners, scanning parameters, and subject cohorts, *etc* [5].

As a promising solution to tackle the cross-domain distribution differences and avoid manual annotations, existing unsupervised domain adaptation (UDA) methods are extremely limited in the selected task. **First,** UDA methods in medical image analysis [6, 7, 8, 9] assume that the images of the target site have the same distribution (one-domain target site). However, in the selected task, the target site has a variety of distributions (multiple-domain target site), limiting their performance on the target site. **Second,** other UDA studies that consider the existence of the multiple domains in the target site [10, 11] are designed specifically for natural image analysis. However, there are many notable differences between medical images and natural images including low contrast, unclear boundaries, *etc.,* which increases the risk of degraded segmentation performance.

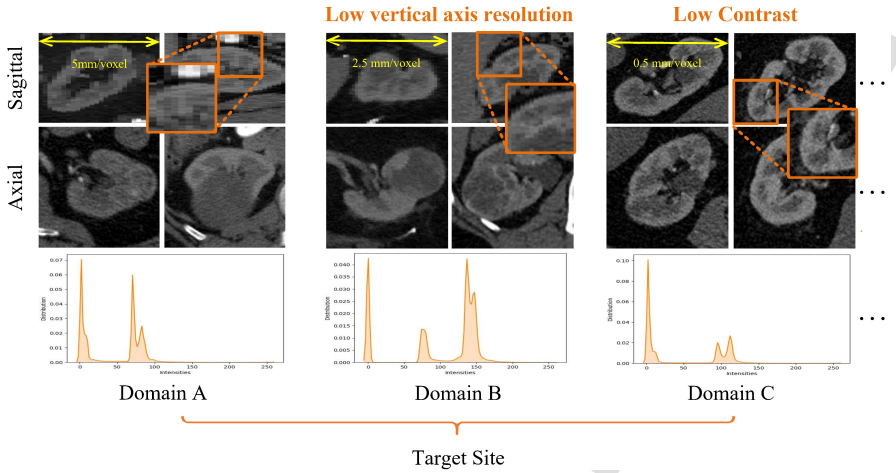

Figure 2: Challenge 1: Multiple domains in a target site result in the distribution differences among multiple domains(among-target differences). Image slices from three different domains within the target site (top) and corresponding intensity distribution (bottom). Domain A, Domain B, and Domain C in the target site show different vertical axis resolutions and different contrasts.

Inherent challenges of the one-to-multiple medical image segmentation based on UDA limit its clinical application. **Challenge 1:** Distribution differences among multiple target domains (among-target differences). As illustrated in Fig. 2, the target site has multiple domains caused by different scanning parameters, protocols, and protocols of image acquisition [4, 5], thus leading to among-target differences. Among-target differences make the data distribution compact at the source site but sparse at the target site. Therefore, merely reducing the distribution difference between the source site and the target site leads to marked differences in model performance on the multiple target domains. For example, the model will perform better on the data of the target domain whose distribution is easier to fit. **Challenge 2:** Distribution differences between one source domain and multiple target domains (source-target differences). As shown in Fig. 3, different angiograph periods make noticeable differences in the distribution of lesions between the data of the one source domain and the multiple target domains. These differences include different grayscale features, different boundary characteristics, and so on, thus causing source-target differences. When applied to the multiple-domain target site, the model will be disturbed by distributions of the multiple target domains that it has never learned from the one source domain, limiting its ability to adapt to these unlearned distributions. Therefore, the segmentation model will have a severe performance degradation on the target site.

For Challenge 1, Pan *et al.* [10] used relatively similar data in the target site to expand the generalization boundary of the model through self-supervision. Cui *et al.* [12] used multi-task learning to train in-domain models tailored for each specific domain and a general-domain model shared by different domains, thus exploiting domain knowledge better by tuning the parameters of these models jointly. **However**, their

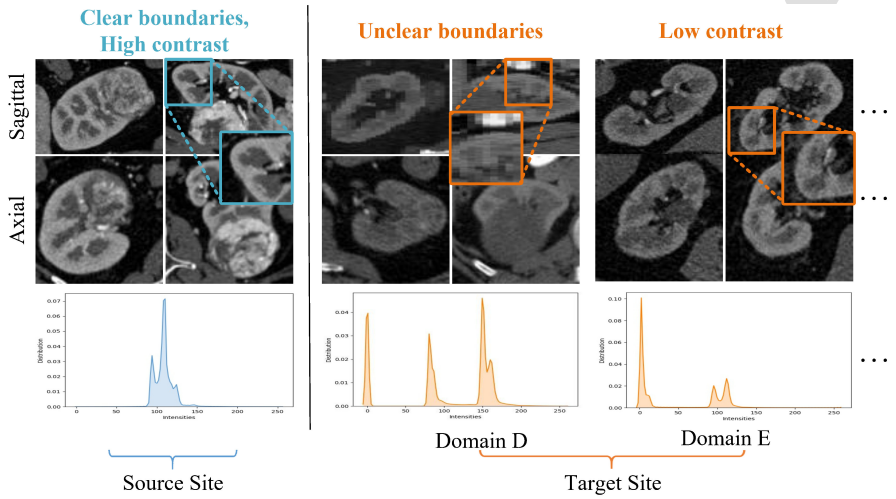

Figure 3: Challenge 2: There are noticeable differences in the distribution of lesions between one source domain and multiple target domains (source-target differences). Image slices from the source site and two domains in the target site (top) and corresponding intensity distribution (bottom). a) The boundary with the medulla in the source site is clearer with vascular highlights, while the boundary with the medulla in Domain D of the target site is unclear. b) Images of the source site have high contrast while images in Domain E of the target site have low contrast. c) The intensity distribution of the source site is markedly different from that of the target site.

performance on the target site is limited because they use the target site data indiscriminately during the training process. It causes the feature learned by the model to be insufficiently complex, resulting in the underfitting of the model. An underfit model cannot perform well on the multiple-domain target site due to the high complexity of the data.

Innovation 1: We propose the *dynamic credible sample strategy (DCSS)* to reduce the among-target differences. As shown in Fig.4, the proposed DCSS gradually expands the generalization boundary of the model by iteratively updating the credible set of the target site data, thus gradually improving the model's adaptation ability. The proposed DCSS has two key elements: **1)** Weighted credibility entropy Evaluation. Based on the ability of entropy to reflect information [13, 14], we use entropy as an evaluation indicator of image credibility and propose a weighted entropy calculation method that assigns different weights to different classes of the foreground rather than a global mean entropy calculation method [10, 15]. It improves the model's attention to some small but important regions such as lesions in the process of calculating credibility, thus better representing the credibility of images. **2)** Domain swell uses a dynamic parameter to dynamically divide the target site data into a credible set and an incredible set based on the proposed weighted credibility entropy evaluation. Increasing the dynamic parameter gradually allows more credible data to participate in model training iteratively, so that the model can learn more complex features, avoiding underfitting and expanding its generalization boundary to cover the data of a multiple-domain target

site.

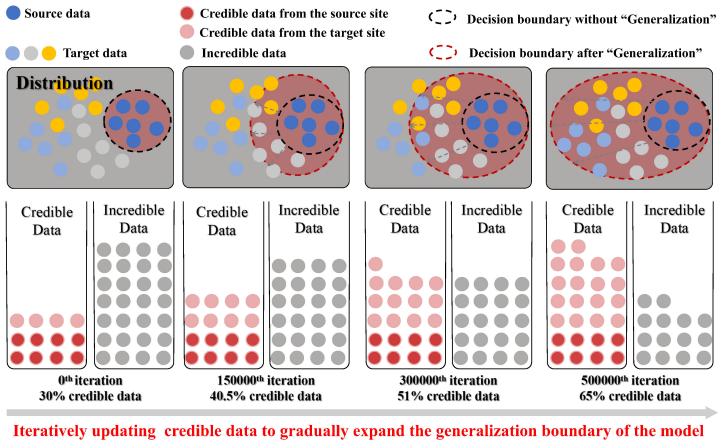

Figure 4: The proposed DCSS gradually expands the generalization boundary of the model by iteratively updating the credible set of the target site data. Our DCSS gradually expands the generalization boundary of the model by iteratively updating the credible data from the target site data. a) At the beginning of DCSS, the credible data consist of data from both the source site and the target site whose weighted entropy is lower. With the increase in the number of training iterations, the proportion of credible data gradually increases which allows the set of credible data is able to include more credible data from the target site, thus iteratively updating the credible data of the target site. b) With the updating of the credible data, the generalization boundary of the model gradually expands to differentiate more data, improving the model's adaptation ability.

For Challenge 2, common ideas are measuring and minimizing the distribution distance between source and target domains to reduce the source-target differences, thus improving the model performance. Based on this idea, previous methods use uncertainty [16, 17, 18] to represent the distance and minimize it through entropy [19]. **However**, they only focus on uncertainty reduction at one level (the voxel level or the domain level), producing the following two limitations: **First,** overfitting of the model is produced because only focusing on voxel-level entropy minimization results in the entropy reduction of the high-entropy regions (e.g., boundary) in the source site. Features from these high-entropy regions will be regarded as learnable features mistakenly by the model and participate in the model training during the adaptation, thus leading to the overfitting of the model [20, 21]. **Second,** the segmentation performance limitation occurs because merely reducing the domain-level uncertainty causes the model to ignore details of segmentations at the output level, thus resulting in poor segmentation performance of the model.

Innovation 2: We propose *hybrid uncertainty learning (HUL)* which reduces the voxel-level and domain-level uncertainty simultaneously, thus minimizing the source-target distribution differences from the detailed and entire perspectives concurrently. The proposed HUL contains the following two complementary methods: **1)** Voxel-

level uncertainty reduction which reduces voxel-level uncertainty by minimizing entropy loss which reduces the high entropy of some incorrect segmentations. Therefore, the segmentation model can pay attention to detailed source-target distribution differences. **2)** Domain-level uncertainty reduction uses adversarial learning to achieve domain adaptation in a global view. Therefore, the segmentation model can pay attention to the distribution differences from the perspective of the entire, reducing source-target differences in a global view.

In this study, we propose an O2M-UDA framework, called dynamic domain adaptation (DyDA), for one-to-multiple medical image segmentation. The contributions of this study include:

- To the best of our knowledge, we develop a novel method that achieves one-to-multiple medical image segmentation for the first time which will reduce the difficulty of training data collection and improve the model's application range. We review the clinical value of this task which will provide a valuable reference for follow-up studies.
- We propose the novel framework *Dynamic Domain Adaptation* (DyDA) and markedly improve the adaptation ability of the model trained on a labeled one-domain source site to an unlabeled multiple-domain target site.
- We propose a novel dynamic strategy, called the *dynamic credible samples strategy (DCSS)* to minimize the among-target differences that iteratively divides the credible set and updates credible samples, thus iteratively expanding the generalization boundary of the model to cover data of the multiple-domain target site and improving the model's adaptation ability.
- We propose a novel learning strategy, called *hybrid uncertainty learning (HUL)* to minimize the source-target differences by reducing the voxel-level and domain-level uncertainty simultaneously, thus allowing the model to minimize the source-target distribution differences from the detailed and entire perspectives concurrently.

2. Related Works

2.1. Unsupervised Domain Adaptation (UDA) in Medical Image Segmentation

Domain shift has been a long-standing problem in medical image segmentation due to the common interscanner or cross-modality variations [22, 23]. The goal of unsupervised domain adaptation is to align the distribution shift between labeled source and unlabeled target data. Existing UDA methods for segmentation can be divided into the following three categories: **First,** in the adversarial learning category, to reduce the cross-domain discrepancy, numerous UDA methods [6, 8, 9, 7] focus on distribution consistency by introducing adversarial learning. **Second,** in the image-to-image translation category, a category of UDA methods have been inspired by image-to-image translation [24] and generates target images conditioned on source data [25, 26]. **Third,** self-supervision with pseudo labels is proposed to generate pseudo labels of target-site data and retrain the model through these pseudo labels, which is relatively simple but

efficient in addressing UDA problems [27, 28]. **However**, these UDA methods for segmentation assume that the images of the target site have the same distribution, ignoring the among-target differences. Therefore, these methods cannot expand the generalization boundary to multiple domains of the target site, thus limiting the model's performance on some target domains.

The proposed DCSS divides the target site into the credible set and the incredible set and then iteratively updates the credible set, thus gradually expanding the model's generalization boundary and improving the model's generalization ability.

2.2. One-to-multiple Unsupervised Domain Adaptation in One-to-multiple Medical Image Segmentation

Currently, few studies have investigated one-to-multiple unsupervised domain adaptation in the field of medical image segmentation [5]. In the field of natural images, previous studies of O2M-UDA break this task into several steps or subtasks. Dai *et al.* [29] subdivided the domain difference between the source domain and the target domain into several smaller domain differences that are easier to minimize. Pan *et al.* [10] proposed a two-step method to separate the target-site data into two parts and then reduce among-target differences by adversarial learning after the model fits [10] **However**, these adaptation methods are limited in the selected task because medical images have different distributions from those of natural images. The distribution of medical images is regarded as a positive distribution while that of natural images is regarded as a Gaussian distribution. Therefore, during training, these methods focus only on the result of the adaptation rather than the process of the adaptation, thus making the model affected by the data whose distribution distance is far. Such an effect will introduce noise into the model, leading to instability in the segmentation model.

The proposed DCSS iteratively updates the credible images of a target site with a dynamic parameter based on the proposed weighted credibility entropy evaluation. Therefore, the generalization boundary of the model will be iteratively expanded with the increase in training iterations, improving the model's adaptation ability.

2.3. Adversarial Learning for Unsupervised Domain Adaptation

Based on a generative adversarial network [30], adversarial-based UDA approaches have shown strong capabilities in learning domain invariant features, even for complex tasks such as semantic segmentation. An adversarial network involves two networks: one network generates the prediction of the segmentation maps for the input images, which could be from the source or the target domain; another network functions as a discriminator to predict the domain labels. The generator network tries to fool the discriminator, thus aligning the distribution shift between the two domains. Previous studies have attempted to align domain shift from different perspectives, including image-level alignment, feature-level alignment, and output-level alignment. **First,** image-level alignment methods transform the source images to appear similar to the target ones or vice versa [31, 25]. For example, CycleGAN [32] performs well in unpaired image-to-image transformation. CycleGAN was also applied in [25] to build generative images for domain alignment. **Second,** feature-level alignment methods extract domain-invariant features of deep neural networks in an adversarial learning scenario to achieve alignment. [33, 34] used a discriminator directly in the feature space

to differentiate the features across domains. Recent studies propose to project the high-dimension feature space to other compact spaces, such as the semantic prediction space [35] or the image space [36]. **Third,** output-level alignment methods propose efficient domain adaptation algorithms through adversarial learning in the output space. For example, [37] proposes an end-to-end model involving structural output alignment for distribution shift. **However**, these methods only use adversarial loss and ignore the voxel-level alignment which makes the model ignore details of segmentations while adapting, thus leading to the model's segmentation performance degradation on the multiple-domain target site.

Based on preserving domain uncertainty minimization, the proposed HUL adds voxel-level uncertainty minimization concurrently. Therefore, the segmentation model minimizes the distribution differences at the level of the entire dataset in a global view and pays attention to details when adapting.

2.4. Uncertainty via Entropy

Uncertainty measurement has a strong connection with unsupervised domain adaptation. Vu *et al.* [15] propose minimizing the target entropy value of the model outputs directly to minimize the distribution differences between the source domain and the target domain for segmentation. Additionally, the entropy of the model outputs is used as a confidence measurement for transferring samples across domains [18, 38].

The proposed HUL minimizes the entropy to reduce the uncertainty at the voxel level and the domain level through the cross entropy and the adversarial entropy, thus minimizing the source-target differences from the detailed and the entire perspective simultaneously.

2.5. Medical Image Segmentation

Medical image segmentation has been considered the most essential medical imaging process because it extracts the region of interest (ROI) which are organs or lesions, through a semiautomatic or automatic process [39]. Existing medical image segmentation methods can be divided into the following two types according to whether they use convolutional neural networks (CNNs): traditional works and CNN-based methods. **1)** Traditional methods perform segmentation based on the characteristics of the medical images, including threshold-based [40], clustering-based [41, 42], and region-based works [43, 44]. **2)** CNN-based methods have been promising alternatives for traditional segmentation methods striking benefits from the remarkable success of deep learning in computer vision. The most popular structure is the encoder-decoder structure, which has an encoder processing and a decoder processing [45]. During the encoder processing, the image content is encoded by multiple convolutional layers from low to high levels. In the decoder stage, the prediction mask is obtained by multiple upsampling (uppooling or deconvolutional) layers. Based on the proposed encoder-decoder architectures, numerous segmentation methods are proposed to improve the proposed encoder-decoder architectures, which include the design of the network backbone [46, 47], network function block [48], loss function [49], feature representation [50, 51]. **However,** traditional methods are strongly affected by the image intensity or texture information, and CNN-based methods heavily rely on the annotations of

medical images that are difficult to obtain and require expert knowledge, limiting their application in the one-to-multiple medical image segmentation task.

The proposed DyDA is a one-to-multiple medical image segmentation framework that requires no additional annotations for images of the multiple-domain target site.

3. Methodology

As shown in Fig. 5, the proposed DyDA minimizes the among-target differences by gradually expanding the model's generalization boundary and the source-target differences by simultaneously minimizing the differences from the detail and entire perspective, thus achieving unsupervised domain adaptation from a one-domain source site to a multiple-domain target site. This process provides notable innovations in two collaborative components: **1)** the proposed Dynamic Credible Sample Strategy (DCSS, Sec. 3.2, Fig. 5 a)) gradually expands the model's generalization boundary and improves the model's adaptation ability, thus minimizing the among-target differences. DCSS uses the weighted entropy value to increase attention to some small but important regions, thus better representing the credibility of target-site images. Based on the target-site image's credibility and a dynamical ratio, the proposed DCSS iteratively updates the credible set divided from the target site, thus gradually expanding the model's generalization boundary. **2)** the proposed hybrid uncertainty learning (HUL, Sec. 3.3, Fig. 5 b)) minimizes the source-target differences from the detail and entire perspective by reducing the voxel-level and domain-level uncertainty simultaneously. HUL reduces the uncertainty at both levels by minimizing two different entropy losses.

3.1. Problem Formulation

A labeled source site dataset $D_s = \{(x_s, y_s)|x_s \in \mathbb{R}^{H\times W\times D\times 1}, y_s \in (1, C)^{H\times W\times D}\}$ and an unlabeled target site dataset $D_t = \{x_t|x_t \in \mathbb{R}^{H\times W\times D\times 1}\}$ are given, where x_s and x_t correspond to the source and target site 3D images, respectively; y_s is the label for the corresponding source image which provides the annotation of the voxel at (H, W, D) in the form of a one-hot vector; and H, W, and D are the height, width, and depth of the images, respectively. There exist n_s image-label pairs in D_s and n_t images in D_t. The problem can be formulated as follows:

$$\min_{\theta_G} \begin{cases} \frac{1}{n_s}\sum_{x_s} \mathcal{L}_S(G(x_s), y_s), \\ \frac{1}{n_t}\sum_{x_t} \mathcal{L}_T(G(x_t)), \end{cases} \tag{1}$$

where $\mathcal{L}_S$ is the loss of the source site image and $\mathcal{L}_T$ is the loss of the target site image.

The proposed DCSS is used to minimize the among-target differences. Considering that the data distribution in medical images presents an increase in data-style similarity, the proposed DCSS divides the images into the credible set D_c and the incredible set D_i based on the proposed Weighted Credibility Entropy Evaluation, and then uses a dynamic parameter $\lambda(r)$ to iteratively update the credible set to reduce the among-target differences. We denote n_c and n_i to represent the number of credible samples and incredible samples, respectively. In the experiment of this study, we use the source site images to initialize the credible set. However, because no label is provided for the target site images, we regard the segmentation map for x_i generated by the segmentation

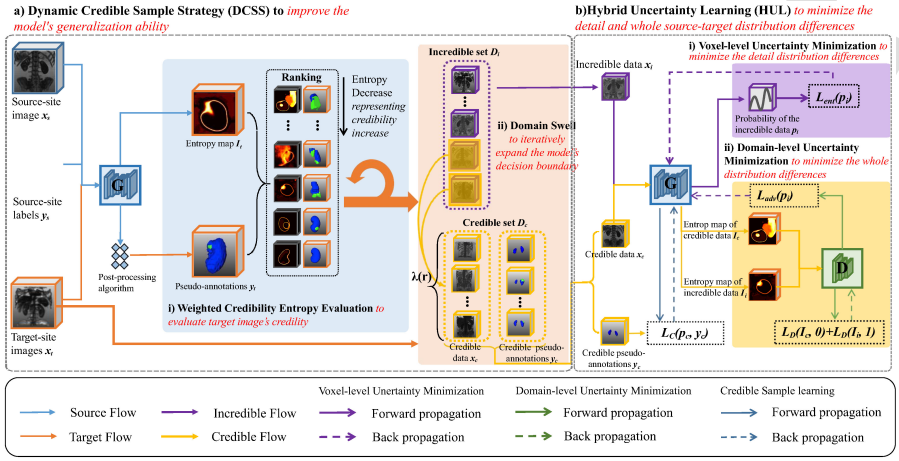

Figure 5: The proposed DyDA for the one-to-multiple medical image segmentation task contains the hybrid uncertainty learning module (HUL) and the dynamic credible sample strategy (DCSS) module. In a), the proposed DCSS extracts credible samples and incredible samples based on the proposed weighted credibility entropy evaluation and a dynamic division ratio. By iteratively increasing the ratio, the credible set is gradually updated, thus gradually expanding the generalization boundary from near to far and improving the segmentation model's adaptation ability on a multiple-domain target site. In b), the segmentation model takes the divided credible set x_c, its pseudo-annotation y_c, and the divided incredible set x_i as input. The proposed HUL reduces the voxel-level and domain-level uncertainty simultaneously to minimize the source-target differences from the detail and entire perspective.

model G after post-processing as its pseudo-label $p_i = G(x_i)$. We define y_c as the pseudo-annotation of x_c which is its label or pseudo-label. Specifically, when x_c is the source-site image, y_c denotes the given corresponding labels of $x_c = x_s$, that is, $y_c = y_s$. When x_c is the target site image divided into the credible set, y_c represents the one-hot encoded segmentation map $y_c = G_{DCSS}(x_c)$ which is regarded as the pseudo-label of x_c in HUL:

$$y_c = \begin{cases} y_s, & if\ x_c = x_s \\ G_{DCSS}(x_c), & if\ x_c \in D_t. \end{cases} \tag{2}$$

The segmentation network G_{DCSS} in the proposed DCSS consists of the trained segmentation model G and the post-processing algorithm for the segmentation results:

$$G_{DCSS}(x_c) = G(x_c) + F_{post}(G(x_c)), \tag{3}$$

where $F_{post}(\cdot)$ is the post-processing algorithm for pseudo-labels. In the experiment, we use the largest connected domain algorithm as the post-processing algorithm to eliminate part of the false-positive segmentation results.

The proposed HUL takes the credible sample x_c, its corresponding pseudo-annotation y_c, and the incredible sample x_i as input. For the credible sample x_c, the segmentation

model G will generate its predicted segmentation map $p_c = G(x_c)$ and is optimized by $\mathcal{L}_C$. For the incredible part, G is optimized by $\mathcal{L}_I$ which is the combination of $\mathcal{L}_{ent}$ and $\mathcal{L}_{adv}$ denoting the loss of the minimization entropy and adversarial entropy, respectively. Therefore, the overall segmentation model G is optimized by $\mathcal{L}_{HUL}$, which is the optimization goal in the proposed scenario:

$$\min_{\theta_G} \left\{ \begin{array}{l} \frac{1}{\lambda(r)\times n_t} \sum_{x_c} \mathcal{L}_C(p_c, y_c) + \\ \frac{1}{(1-\lambda(r))\times n_t} \sum_{x_i} \mathcal{L}_I(p_i), \end{array} \right. \tag{4}$$

where $\lambda(r)$ is the value of the dynamic ratio parameter in the $r-th$ round. $\lambda(r)$ is defined in Eq. 6 and its detailed calculation is presented in Eq. 7.

3.2. Dynamic credible sample strategy (DCSS) to improve the model's adaptation ability

The proposed DCSS iteratively updates the divided credible samples using a dynamic ratio to gradually expand the generalization boundary of the segmentation model, thus minimizing the among-target differences and improving the adaptation ability of the model.

1) Weighted credibility entropy evaluation to evaluate the target image's credibility:

Inspired by entropy being able to reflect information [13], we used the entropy value of the entropy map generated by the predictions to evaluate the target image's credibility. We have observed that the model trained with the image from the source site tends to produce overconfident predictions on source-like images whose data distributions are similar to the source site. The overconfident predictions of the target image have lower entropy values, while the underconfident predictions have higher entropy values. Therefore, we can determine whether the prediction is from a source-like image or a target-like image based on its entropy value. A lower entropy value means a more source-like image that has a lower credibility level in the target site.

Existing methods [10, 15] have applied the global mean entropy to evaluate the credibility levels of the target site images. However, when the regions to be segmented only account for some part of the image, such as the lesion part, their high entropy will have little effect on the global entropy, thus leading to the neglect of these regions in the credibility evaluation. Such neglect causes the credible-set division to focus more on the large but unimportant regions in the image, which leads to segmentation performance degradation for some tiny structures such as the lesions in medical images of the target site.

The proposed Weighted Credibility Entropy Evaluation pays more attention to small but important structures such as lesions in the credibility evaluation for target site images. By assigning different weights to different structures, the weighted entropy results can improve the influence of small but important structural regions when calculating credibility. Specifically, this method assigns different weights to each class of the foreground and calculates the weighted entropy R_t of x_t in the following way:

$$R_t = \sum_c^C \omega_c \times \frac{1}{N(F_{post}(p_t)^{(h,w,d,c)})} \sum_{h,w,d} I_t^{(h,w,d,c)}, \tag{5}$$

where $\omega_c \in (0, 1]$ is the weight assigned to class c of the foreground; I_t is the entropy map of the pseudo-label p_t for the target data x_t generated by the proposed segmentation model; and $N(F_{post}(p_t)^{(h,w,d,c)})$ is the number of voxel regions predicted to be class C in p_t after post-processing. The ratio of different values for ω is inversely proportional to the ratio of voxels for different classes in the source images.

2) Domain Swell to iteratively expand the generalization boundary of the model: the proposed Domain Swell gradually expands the generalization boundary of the model by iteratively updating the credible set, thus gradually reducing the among-target differences. Because the trained models tend to produce lower entropy prediction for the images that are more similar to a source site [15], the proposed Domain Swell divides the target site images with low weighted entropy values into the credible set and the rest into the incredible set. Each division is based on the dynamic ratio $\lambda(r)$, which represents the ratio of the credible samples to the target-site images in the $r-th$ round. $\lambda(r)$ is defined in the following way:

$$\lambda(r) = \frac{n_c(r)}{n_t}, \tag{6}$$

where $n_c(r)$ and n_t are the number of credible samples from the target site in the r-th round and the number of target-site images, respectively.

By iteratively executing the division, the dynamic ratio will expand from the initial ratio to the final ratio in the following way:

$$\lambda(r) = \frac{r}{\gamma} \times (\lambda_{final} - \lambda_{init}) + \lambda_{init}, \tag{7}$$

where λ_{init} and λ_{final} are two hyperparameters representing the initial and final ratios of credible samples, respectively; γ is the total rounds of training DCSS; and r is the $r - th$ division round in DCSS. The value of r is between 1 and the total rounds of training DCSS, which means $r \in [1, \gamma]$.

3.3. Hybrid uncertainty learning (HUL) to minimize the detail and entire source-target distribution differences

The proposed HUL reduces the voxel-level and the domain-level uncertainty simultaneously to minimize the source-target distribution differences from the detail and entire perspective concurrently.

1) Voxel-level Uncertainty Reduction to minimize the detail distribution differences: While learning for the incredible samples, the segmentation model G takes the incredible samples x_i as input and generates a segmentation map $p_i = G(x_i)$. For the voxel at (h, w, d), the value of each dimension after normalization by the softmax layer is a probability $\in [0, 1]$. The Shannon entropy[13] calculated on this probability represents the certainty of G in this voxel. By directly reducing the entropy loss $\mathcal{L}_{ent}$, the model G can pay more attention to detailed source-target distribution differences when adapting to the target site. $\mathcal{L}_{ent}$ is the first part of the incredible image loss:

$$\mathcal{L}_{ent}(p_i) = -\sum_c^C \sum_{h,w,d} p_i^{(h,w,d,c)} \log(p_i^{(h,w,d,c)}). \tag{8}$$

2) Domain-level Uncertainty Reduction to minimize the entire distribution differences: Minimizing $\mathcal{L}_{ent}$ alone takes the generalization boundary of the model to the

low-density area at a voxel-wise level. To manage the domain-level distribution differences of the entire dataset, the proposed HUL introduces adversarial learning at the output level to unify the distribution. Adversarial learning consists of two networks: the generator network G_g and the discriminator network D. In this study, the generator network takes the incredible image x_i as input and produces its pseudo-label $p_i = G(x_i)$. Therefore, the generator network G_g is also the segmentation network G. Then, D takes the entropy map I_i as input and produces its classification results $\in \{0, 1\}$ to judge whether the input image comes from the credible set or the incredible set, where I_i is the entropy map of p_i after entropy calculation $I_i^{(h,w,d)} = -p_i^{(h,w,d)} \cdot \log p_i^{(h,w,d)}$. G is optimized by reducing $\mathcal{L}_{adv}$:

$$\begin{aligned} \mathcal{L}_{adv}(p_i) &= \frac{1}{n_i} \sum_{x_i} \mathcal{L}_D(I_i, 0) \\ &= \frac{1}{n_i} \sum_{x_i} \mathcal{L}_D(-p_i^{(h,w,d)} \cdot \log p_i^{(h,w,d)}, 0), \end{aligned} \tag{9}$$

where n_i is the number of the incredible samples.

The parameter θ_D of D is optimized by minimizing $\mathcal{L}_D$ which is the cross-entropy classification loss:

$$\min_{\theta_D} \mathcal{L}_D = \frac{1}{n_c} \sum_{x_c} \mathcal{L}_D(I_c, 0) + \frac{1}{n_i} \sum_{x_i} \mathcal{L}_D(I_i, 1), \tag{10}$$

where n_c and n_i are the number of credible samples and the incredible samples, respectively.

Combining Eq.8 and Eq.9, the incredible set loss $\mathcal{L}_I$ is as follows:

$$\mathcal{L}_I(p_i) = \eta_{ent} \mathcal{L}_{ent}(p_i) + \eta_{adv} \mathcal{L}_{adv}(p_i), \tag{11}$$

where η_{ent} and η_{adv} are two hyperparameters that weigh the proportion of the voxel-level and domain-level uncertainty reduction, respectively. In this study, these parameters are set by experience.

3.4. The Details of the proposed framework:

While learning for credible samples, the segmentation network G takes the credible sample x_c as input and generates a predicted segmentation map $p_c = G(x_c)$ with C dimensions. Every voxel of p_c will become a discrete distribution in each dimension after the softmax layer. The proposed approach to credible sample learning follows the famous method [46], using soft dice loss and cross-entropy loss concurrently. In the case of the pseudo-annotation y_c, G is optimized by minimizing the sum of the two losses:

$$\mathcal{L}_C(p_c, y_c) = \mathcal{L}_{CE}(p_c, y_c) + \mathcal{L}_{dice}(p_c, y_c), \tag{12}$$

where $\mathcal{L}_{CE}$ is cross-entropy loss and $\mathcal{L}_{dice}$ is soft dice loss, which are defined as follows:

$$\begin{aligned} \mathcal{L}_{dice}(p_c, y_c) &= -\frac{2}{|C|} \sum_{c \in C} \frac{\sum_{h,w,d} p_c \times y_c}{\sum_{h,w,d} p_c + \sum_{h,w,d} y_c}. \\ \mathcal{L}_{CE}(p_c, y_c) &= -\sum_c^C \sum_{h,w,d} y_c^{(h,w,d,c)} \log(p_c^{(h,w,d,c)}). \end{aligned} \tag{13}$$

4. Experimental Configurations

4.1. Dataset

We evaluate the proposed DyDA framework on two important one-to-multiple medical image segmentation tasks with different image modalities:

1) Kidney&Tumor Segmentation on JSPH → KiTS19 evaluates the proposed DyDA in kidney&tumor segmentation of abdominal CT images using JSPH as the one-domain source dataset and KiTS19 as the multiple-domain target site dataset. **One-domain source site JSPH** is a collection of abdominal CT scans from Jiangsu People's Hospital. All images were acquired by a Siemens dual-source 64-slice CT scanner. The contrast media was injected during all the CT image acquisition and all the images were scanned in the early-artery phase. The slices of all the images are fixed at 0.5mm with annotation. Because all the images were collected from the same hospital and there was no difference in acquisition parameters and equipment, the data distribution of these images varies little, making JSPH a one-domain dataset. **The multiple-domain target site KiTS19** [52] is a segmented CT image dataset collected from more than 60 medical centers and scanned at different phases. Manual segmentation results of training data are provided for the kidneys and tumors in these images. The KiTS19 test set consists of 90 CT scans and the JSPH test set consists of 70 CT scans. The test set results of KiTS19 are uploaded for online evaluation.

2) RV&Myo&LV Segmentation on ACDC → MyoEmidec evaluates the proposed DyDA in RV&Myo&LV segmentation of cardiac magnetic resonance (CMR) images using ACDC as the one-domain source dataset and MyoEmidec as the multiple-domain target site dataset. **One-domain source site ACDC** [53] was provided by MICCAI'17 Automatic Cardiac Diagnosis Challenge. This dataset consists of shot-axis cardiac cine-MRIs of 100 patients for training, and of 50 patients for testing. Manual segmentation results of training data are provided for the right ventricle (RV), left ventricle (LV), and Myo during the end-diastolic and end-systolic phases (Myo). **The multiple-domain target site MyoEmidec** is composed of **MyoPS2020** [54] and **EMIDEC** [55]. This dataset consists of 375 CMR images from **five** domains for training. MyoPS2020 provides 225 CMR images from three domains: balanced steady-state free precession (bSSFP), late gadolinium enhancement (LGE), and T2-weighted CMR (T2). EMIDEC contains 150 CMR images from two domains: normal MRI after injection of a contrast agent (normal) and myocardial infarction with a hyperenhanced on DE-MRI (pathological). Manual segmentation results are available for all images of the left ventricle (LV), right ventricle (RV), and myocardium (Myo). The test set of MyoEmidec consists of 55 CMR images where 31 images are from EMIDEC and 24 images are from MyoPS2020. The test set results of MyoEmidec were evaluated by comparison with their ground-truth images. No label of any target site image is involved in the training process.

4.2. Implementation details

The segmentation network G in DyDA is optimized by the SGD optimizer with a batch size of 5 and a learning rate of 1×10^{-2}. We use the ordinary 3D UNet [45] to realize G. The entropy loss weight η_{ent} is configured to 0.005 while the adversarial loss weight η_{adv} is 0.0007. In the framework, the discriminator D is optimized by the

Adam optimizer with a learning rate of 1×10^{-3}. According to the proposed experience, the model is trained for a total of 500,000 iterations. After the initial iteration of 250,000, the source site data are regarded as a portion of the credible data to participate in iterative training. The credible sample division is automatically performed every 125,000 iterations and repeated for a total of 2 rounds. λ_{init} is set to 0.3 while λ_{final} is set to 0.8. This framework is implemented on PyTorch and runs on NVIDIA TESLA V100. In the JSPH $\rightarrow$ KiTS19 experiment, random rotation (-20°to +20°), mirror flip, scaling (0.7 to 1.4), and random cropping (128*128*128) are used to augment the image online. To have a better average segmentation performance on the overall medical images, we set the ratio of weights to different organs and lesions as the inverse ratio of their voxels in the source labels of the training dataset. For the convenience of calculation, we expand them to the minimum integral multiple. Specifically, the voxel ratio of tumor and kidney in the JSPH training dataset is 1:3, and the voxel ratio of RV, Myo, and LV in the ACDC training dataset is 3:1:1. Therefore, when calculating the weighted credibility entropy in the JSPH $\rightarrow$ KiTS19 experiment, the weights of the tumor and kidney were initially 0.75 and 0.25, and then expanded to their minimum integral multiple. Finally, we set the weights of the tumor and kidney (ω_{tumor} and ω_{kidney}) to 3 and 1, respectively. Similarly, when calculating the Weighted Credibility Entropy in the ACDC $\rightarrow$ MyoEmidec experiment, we set the weights of RV, Myo, and LV (ω_{RV}, ω_{Myo}, and ω_{LV}) to 1, 3, and 1, respectively.

4.3. Comparison settings

To explain the superiority of the proposed framework, we compared the proposed DyDA with three existing UDA methods on the KiTS19 test set and MyoEmidec test set: MinEnt [56], AdvEnt [15], IntraDA [10]. MinEnt focuses on reducing the voxel-level uncertainty by minimizing entropy loss. AdvEnt focuses on reducing domain-level uncertainty by minimizing adversarial loss. For a fair comparison, nnU-Net [46] is applied as the benchmark for all methods. All these methods were written into 3D for comparison in the proposed 3D datasets.

4.4. Evaluation metrics

In the JSPH $\rightarrow$ KiTS experiment, we follow the evaluation metric of the KiTS19 competition [52] and use the Dice coefficient (Dice) to evaluate all methods.

In the ACDC $\rightarrow$ MyoEmidec experiment, to evaluate all methods, we use dice for pixel-wise accuracy measurement and Hausdorff distance (HD) for boundary agreement assessment following MyoPS2020 [54] and EMIDEC [55].

Dice and HD are calculated in the following ways:

$$Dice(G,P) = \frac{2|G \cap P|}{|G| + |P|} \times 100\% \tag{14}$$

$$HD(G,P) = max(h(G,P), h(P,G)) \tag{15}$$

$$h(G,P) = \max_{g\in G} \min_{p\in P} ||g-p||, \tag{16}$$

where P and G are the predicted result and the true label respectively. A higher Dice and a lower HD indicate better performance.

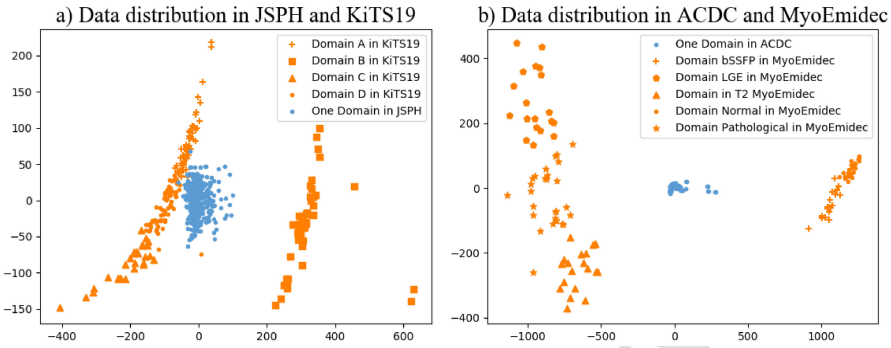

Figure 6: Distribution analysis of the dataset shows the source-target differences and among-target differences in the two one-to-multiple image segmentations. In a), data distribution obtained from features including mean, standard deviation, spacing, and signal-to-noise ratio are shown, which are extracted from JSPH and KiTS19, dimension reduced using PCA[57] and clustered using K-means [58]. Blue points and orange markers in different shapes indicate one domain in JSPH and different domains in KiTS19, respectively. In b), the data distribution after dimension reduction of many features are shown which are extracted from ACDC and MyoEmidec. Blue points, orange plus, orange pentagram, orange triangle, orange point, and orange asterisk indicate the one domain in ACDC, bSSFP, LGE, T2, Normal, and Pathological domain in MyoEmidec, respectively.

5. Results and Analysis

5.1. Dataset Distribution analysis

As shown in Fig. 6, the datasets used in the two one-to-multiple image segmentation experiments have marked source-target differences and among-target differences. The primary features of the data including their mean, standard deviation, spacing, and signal-to-noise ratio are extracted and dimensionally reduced using PCA [57]. As shown in Fig. 6 a), four different data distributions in KiTS19 and one data distribution in JSPH are shown, indicating that there are among-target differences within KiTS19 and source-target differences between JSPH and KiTS19. Although the data distribution differences shown in Fig. 6 b) are not as large as those in Fig. 6a), source-target differences between ACDC and MyoEmidec can be observed. Among-target differences within MyoEmidec are clearly observed.

5.2. Quantitative evaluation for the proposed DyDA

As shown in Tab. 1, the proposed DyDA has achieved excellent performance on two one-to-multiple medical image segmentation tasks. As shown in Tab. 1 a) and Tab. 1 b), the average dice on the KiTS19 test and MyoEmidec test set is 83.8% and 48.1%, respectively, which has improved by 21.7% and 9.2%, respectively, compared with that before UDA. These results indicate that DyDA can improve the performance of the segmentation model in different one-to-multiple medical image segmentation tasks. Compared with the three existing UDA methods, the proposed HUL and the

Table 1: The quantitative evaluation demonstrates the advantages of the proposed DyDA on two one-to-multiple medical image segmentation tasks.

(a) KiTS19 test set[52] (multiple-domain target site)

Method	Kidney dice (%)	Tumor dice(%)	Average(%)
Before UDA [46]	81.3	42.8	62.1
MinEnt [56]	87.3(+5.9)	65.7(+22.9)	76.4(+14.4)
AdvEnt [15]	86.9 (+5.6)	67.4(+24.6)	77.1(+15.1)
IntraDA(λ=0.5) [10]	85.4(+4.1)	39.2(-3.6)	62.3(+0.2)
IntraDA(λ=0.7) [10]	87.7(+6.3)	60.2(+17.4)	73.9(+11.9)
Proposed HUL	91.2(+9.9)	69.9(+27.1)	80.6(+18.5)
Proposed DyDA	**92.3(+11.0)**	**75.4(+32.6)**	**83.8(+21.7)**

(b) MyoEmidec test set[54][55](multiple-domain target site)

	RV		Myo		LV		Average	
Method	Dice (%)	HD(mm)	Dice (%)	HD(mm)	Dice (%)	HD(mm)	Dice (%)	HD(mm)
Before UDA [46]	24.2	44.6	63.5	21.2	28.9	32.9	38.9	32.9
MinEnt [56]	25.2 (+1.0)	31.3 (-13.3)	69.1(+5.6)	8.3(-12.9)	34.4(+5.5)	24.1(-8.8)	42.9(+4.0)	21.2(-11.7)
AdvEnt [15]	24.8(+0.6)	36.5(-8.1)	67.9(+4.4)	7.3(-13.9)	33.1(+4.2)	23.7(-9.2)	42.0(+3.1)	22.5(-10.4)
IntraDA (λ=0.5) [10]	24.9(+0.7)	38.8(-5.8)	65.7(+2.2)	15.5(-5.7)	30.1(+1.2)	29.3(-3.6)	40.2(+1.3)	27.9(-5.0)
IntraDA (λ=0.7) [10]	25.8(+1.6)	33.3(-11.3)	73.1(+9.6)	10.0(-11.2)	35.7(+6.8)	17.9(-15.0)	44.9(+6.0)	20.4(-12.5)
Proposed HUL	26.2 (+2.0)	22.8(-21.8)	74.1(+10.6)	4.7(-16.5)	38.2(+9.3)	16.2(-16.7)	46.2(+7.3)	14.6(-18.3)
Proposed DyDA	**26.8(+2.6)**	**12.0(-32.6)**	**77.4(+13.9)**	**3.1(-18.1)**	**40.1(+11.2)**	**13.5(-19.4)**	**48.1(+9.2)**	**9.6(-23.3)**

proposed DyDA obtain better performance in the two segmentation tasks. As shown in Tab. 1 a), the proposed HUL achieves a dice of 91.2% and 69.9% on kidney and tumor, respectively, which is better than AdvEnt and MinEnt. The proposed HUL obtains an average dice of 80.6% on the kidney&tumor segmentation task, which is 4.1% higher than that of MinEnt and 3.4% higher than that of AdvEnt. Similarly, with the evaluation of the three evaluation metrics, the proposed HUL still achieves better performance on RV, Myo, and LV than MinEnt, AdvEnt, and IntraDA. The performance of RV, Myo, and LV on Dice and HD is 26.2%, 22.8mm, 74.1%, 4.7mm, 38.2%, and 16.2mm, respectively. The proposed HUL obtains 46.2% average dice and 14.6mm average HD in the RV&Myo&LV segmentation task, which are 3.3% and 6.6mm better than those of MinEnt. The proposed HUL performs 4.2% and 7.9mm better than AdvEnt. These improvements on two evaluation metrics on two one-to-multiple medical image segmentation tasks show that compared with focusing on the reduction of one-level uncertainty, reducing the voxel-level uncertainty and the domain-level uncertainty simultaneously can improve the segmentation performance of the model. As shown in Tab. 1 a) the proposed DyDA achieves a 9.8% average dice improvement compared with that of IntraDA($\lambda = 0.7$), which is 83.8% and 73.9% respectively. Similarly, as shown in Tab. 1 b), the proposed DyDA achieves an average 3.8% improvement in dice and 10.8mm reduction in HD compared with those of IntraDA($\lambda = 0.7$). These improvements on the two evaluation metrics on these two on-to-multiple medical image segmentation tasks show that compared with using the target site data indiscriminately, the dynamic division of the credible sample set enables the model to learn more com-

Table 2: Comparison between fully-supervised method and the proposed DyDA on the target site dataset.

KiTS19 test set[52]			
Method	Kidney	Tumor	AVG
Before UDA [46]	81.3	42.8	62.1
Proposed DyDA	**92.3**	**75.4**	**83.8**
KiTS19 No.1(fully-supervised) [46]	97.9	85.4	91.7

plex features, thus improving the adaptation ability of the model.

As illustrated in Tab. 2, the proposed DyDA has also achieved good performance on the multiple-domain target site compared to the fully supervised approach in the JSPH→KiTS experiment, which trains and tests with additional target-site labels. The final average dice is 83.8%, which reaches 90% of the first-place performance (fully supervised method) in the KiTS19 Challenge [52].

5.3. Qualitative evaluation

As shown in Fig. 7, the proposed DyDA has extraordinary visual superiority in different segmentation tasks and different datasets, which will provide visual guidance in clinical surgery. As shown in Fig.7 a) and Fig.7 b), the proposed DyDA obtains a segmentation result close to the ground truth compared with IntraDA both in kidney&tumor segmentation and RV&Myo&LV segmentation. This result likely occurs because the proposed HUL reduces the voxel-level uncertainty and the domain-level uncertainty simultaneously, thus improving the segmentation performance on the target site from the details and the global view. This result also proves that the proposed DyDA is applicable and effective to different segmentation tasks and different datasets for domain adaptation. The proposed DyDA can improve the segmentation model's generalization ability and avoid underfitting by gradually expanding the generalization boundary. In axial plane case 1, axial plane case 2, bSSFP case, and LGE case, the proposed DyDA corrected the under-segmentation and the over-segmentations (red arrows) compared with IntraDA.

5.4. Ablation study

As Tab. 3 illustrates, the proposed innovation has produced notable improvements in performance. The basic U-net segmentation model without UDA achieved 81.3% and 42.8% dice on the kidney and tumor, respectively. When using $\mathcal{L}_{ent}$, the network focuses on reducing the entropy of the target image at the voxel level, achieving 14.3% average dice improvement. When using $\mathcal{L}_{adv}$, which is adversarial training based on the domain level, compared with the $\mathcal{L}_{ent}$ strategy, the tumor part is increased by 1.7%, but the kidney part is reduced by 0.4%. The proposed HUL benefits from both these complementary methods and captures complex cross-domain knowledge from both the voxel level and the domain level, achieving 18.5% average dice improvement, which is better than using either alone. When using DCSS to cope with the among-target differences, the average dice is improved compared to only reducing the source-target differences by 21.7%. Finally, we obtain the best average dice of 83.8% for the proposed one-to-multiple medical image segmentation task.

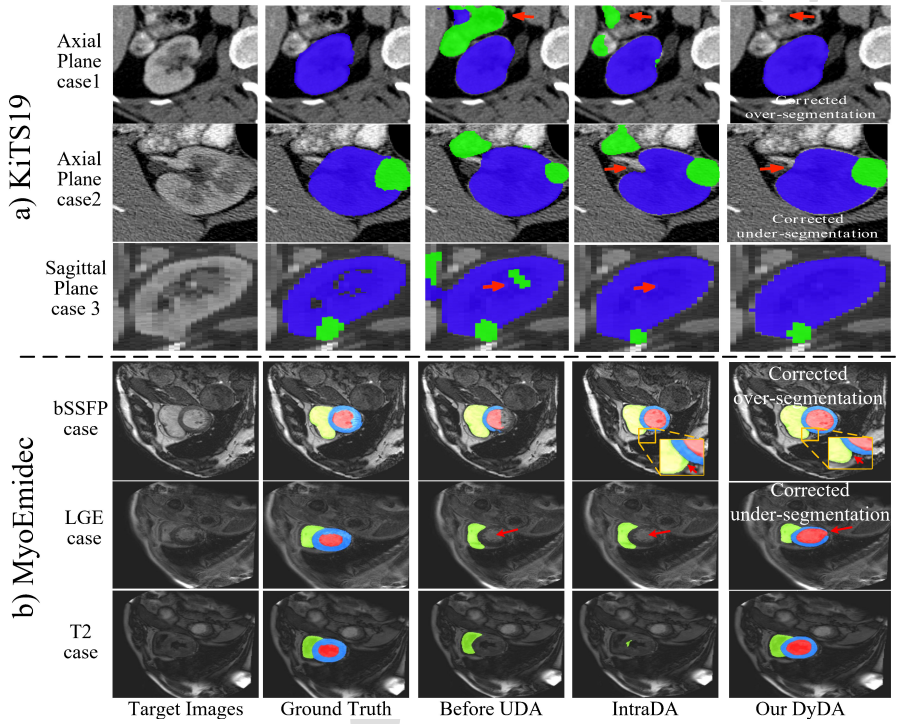

Figure 7: The qualitative evaluation shows the visual superiority of the proposed DyDA on one-to-multiple medical segmentation compared with the existing UDA method and before UDA segmentation. Each row presents one typical case in the target site, from left to right: target images, ground truth, before UDA segmentations, IntraDA segmentation results, and Our DyDA results. The structures of the kidney, tumor, RV, Myo, and LV are indicated by blue, green, green, blue, and red, respectively (best viewed in color) In a), three cases from the axial plane and sagittal plane in KiTS19 are shown. In b), three cases from MyoEmidec in the bSSFP, LGE, and T2 domains are shown.

Table 3: The ablation study analyses the contributions of the proposed innovations on the JSPH → KiTS19 task. The proposed HUL simultaneously reduces the voxel-level uncertainty and the domain-level uncertainty of the model in the target site. The proposed DCSS iteratively extracts credible samples and divides a subset of the target site to iteratively adapt the feature boundaries of the credible data, gradually expanding the generalization boundary of the model and achieving better performance on the target site.

$\mathcal{L}_{ent}$	$\mathcal{L}_{adv}$	DCSS	Dice(%)		
			Kidney	Tumor	AVG
			81.3	42.8	62.1
✓			87.3(+6.0)	65.7(+22.9)	76.4(+14.3)
	✓		86.9(+5.6)	67.4(+24.6)	77.1(+15.0)
✓	✓		91.2(+9.9)	69.9(+27.1)	80.6(+18.5)
✓	✓	✓	**92.3(+11.0)**	**75.4(+32.6)**	**83.8(+21.7)**

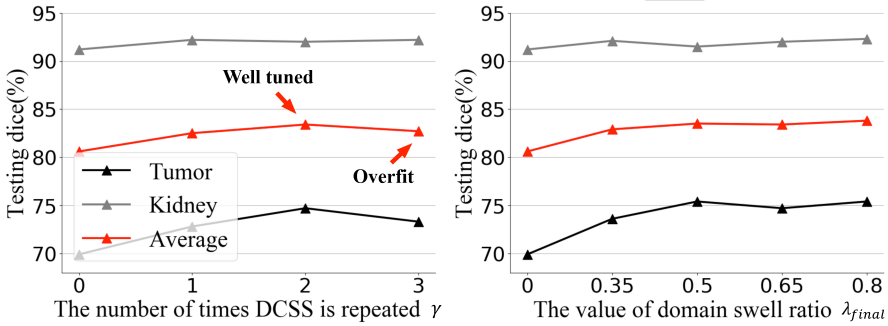

Figure 8: The ablation experiment of γ for DCSS when λ_{final}=0.5 (left) and λ_{final} for the ratio of domain swell when γ=2 (right). The segmentation model performance in the target site shows a trend of rising first and then falling as gamma increases. λ_{final} is positively correlated with the performance first and then remains stable. We use γ=2 and λ_{final}=0.8 as the final hyperparameter combination.

5.5. *Analysis of Hyperparameter*

The most important hyperparameters in DyDA are γ and $\lambda(r)$, where γ represents the rounds of DCSS and $\lambda(r)$ represents the division ratio of the credible set. According to Eq. 7, $\lambda(r)$ is determined by λ_{final}, λ_{init}, and γ jointly. The total training time increases with increasing γ while the iterations in each round remain unchanged. To find the best point between the effect and time, we conduct a study on the γ in the JSPH→KiTS19 experiment. In Fig.8, different γ values are tested for the effect of DyDA on the target site. When γ is 0, no DCSS is dealt with. When γ is less than 2, DCSS is positively correlated with the performance of DyDA. More fine-tuning iterations improved the performance in the target domain. However, when γ is 3, the indicators of the model decrease. We consider the number of training iterations to be too high for a relatively small λ_{final}, thus, the model is overfitted. We consider γ of 2 as the final result, and each round of credible data learning performs 125,000 iterations.

λ_{final} is the ratio of the final credible data in the total target site. A large λ_{final}

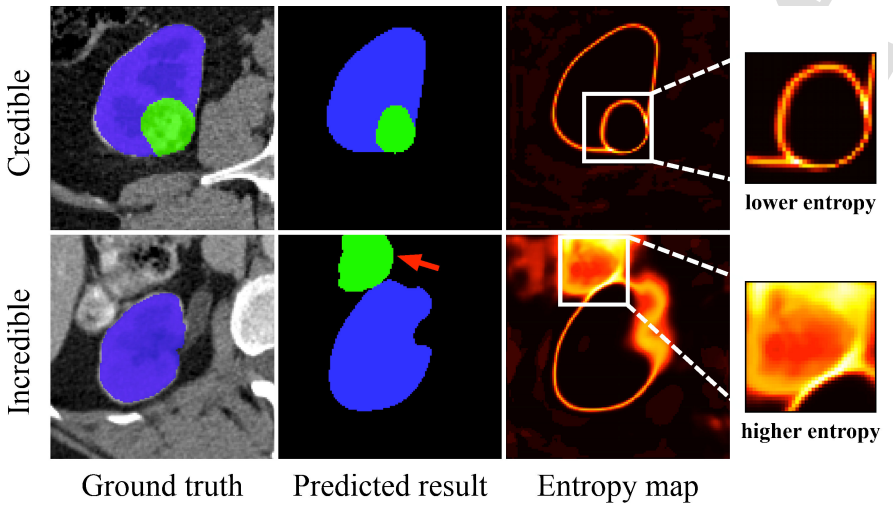

Figure 9: Different regions have different entropy, and higher entropy values generally appear along with incorrect prediction results.

Table 4: The dynamic ratio shows its effectiveness compared with the fixed ratio on the JSPH → KiTS19 task.

KiTS19 test set[52]			
Ratio	Kidney	Tumor	AVG
λ_{init}=λ_{final}=0.3	83.3	66.5	74.9
λ_{init}=λ_{final}=0.8	88.5	70.4	79.5
λ_{init}= 0.3, λ_{final}=0.8	**92.3**	**75.4**	**83.8**

indicates a radical domain swell strategy, however, there is a risk that more false samples will be classified as credible data. To analyze how λ_{final} affects the performance of the segmentation model on the target site, we conduct a study on λ_{final} in the JSPH→KiTS19 experiment. In Fig.8, we evaluate the effect of λ_{final} on the final result when γ is 2 and λ_{init} is 0.3. When λ_{final} is smaller than 0.5, the performance of the segmentation model on the target site is positively correlated with λ_{final}, and when λ_{final} is larger than 0.5, the performance of the segmentation model remains unchanged because when λ_{final} is too small, there are not sufficient target site features to be learned. When λ_{final} is larger than 0.5, sufficient credible data provide the model with sufficiently complex features to learn, and the average dice reaches the highest value of 83.8%.

To analyze the effectiveness of the proposed dynamic ratio, we compare the segmentation performance with the fixed ratio (λ_{final} = λ_{init}). As shown in Tab. 4, the segmentation performance of the dynamic ratio is the best, reaching 92.3% on the kidney and 75.4% on the tumor, respectively.

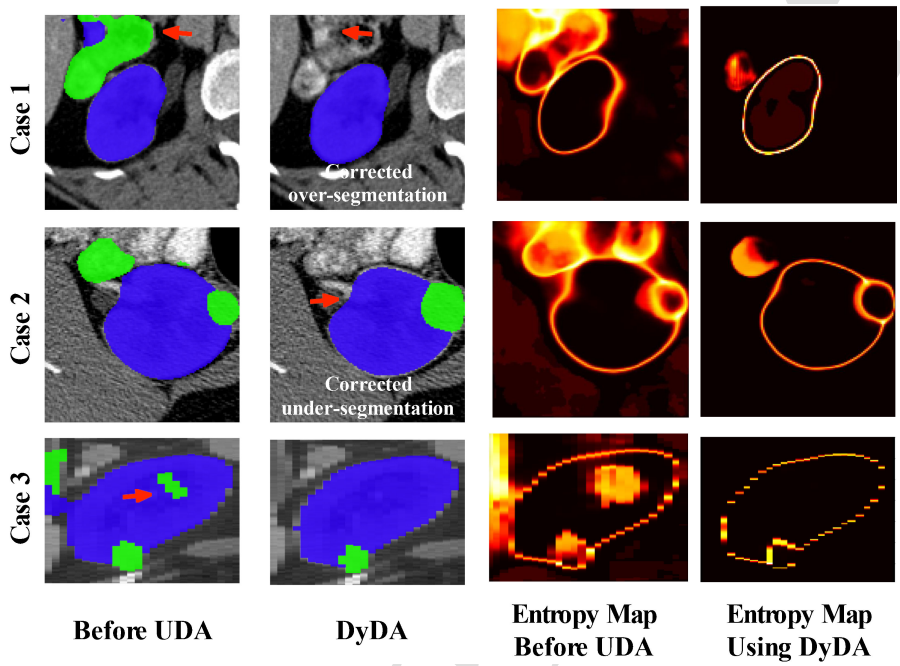

Figure 10: The proposed DyDA obtains accurate segmentations, and the entropy of segmentations after using DyDA is lower.

5.6. Analysis of the credibility for the entropy

1) Credible samples have more accurate predictions. As shown in Fig. 9, comparing the ground truth and prediction results of credible and incredible samples, the external regions that are not the kidney (the area indicated by the red arrow) were mistakenly segmented into tumors in the predicted result of the incredible data while the predicted result of the credible data is consistent with the ground truth.

2) Credible samples have lower entropy. As shown in the third column in Fig. 9, in the credible data, the high-entropy regions such as the boundary in the source site also have low entropy while high entropy is distributed in the mis-segmented regions and the boundaries in the incredible data. Therefore, the overall entropy of the credible data is low, while the entropy of the incredible data is high.

3) Credible samples which have lower entropy contribute to accurate segmentation results. As shown in Fig. 10, the high entropy values of some high-entropy regions have been effectively reduced after using DyDA, and the segmentation results of DyDA are more accurate than those before UDA.

5.7. Analysis of the weighted credibility

As shown in Tab. 5, the weight parameters we selected for the JSPH $\rightarrow$ KiTS19 are optimal. When using the proposed weight parameters (ω_{tumor} = 3, ω_{kidney} = 1), the dice on kidney and tumor have reached 92.3% and 75.4%, respectively which is the

Table 5: The weight parameter study shows the optimality of the proposed selected value on the JSPH → KiTS19 task.

KiTS19 test set[52]			
Weight values	Kidney	Tumor	AVG
ω_{tumor}=1, ω_{kidney}=1	91.9	70.9	81.4
ω_{tumor}=5, ω_{kidney}=1	92.2	72.4	82.2
ω_{tumor}=3, ω_{kidney}=1	**92.3**	**75.4**	**83.8**

best performance of all the results. When the weight of the tumor increased from 1 to 5, the dice on the tumor increased by 1.5%. However, this method still performs worse than the performance of the proposed selected weight parameters, which is 3% lower on the tumor.

6. Discussion and Conclusion

In this paper, we have proposed dynamic domain adaptation (DyDA) for one-to-multiple medical image segmentation. The DyDA minimizes the among-target differences and the source-target differences via the proposed DCSS and HUL. The proposed DCSS iteratively expands the model's generalization boundary by iteratively updating the credible data, thus minimizing the among-target differences and improving the model's adaptation ability. The proposed HUL minimizes the source-target differences from the detail and entire perspectives simultaneously by reducing the voxel-level and domain-level uncertainty concurrently.

Experiments verified that: **1)** DyDA improves the model's segmentation performance on the multiple-domain target site in different datasets and different segmentation tasks. The proposed DyDA is applicable and effective to different tasks and datasets for one-to-multiple medical image segmentation. **2)** Dynamic and iterative division of the credible set can make the generalization boundary of the model gradually expand to cover multiple-domain data, thus reducing the among-target differences. **3)** Simultaneously reducing the voxel-level and domain-level uncertainty improves the segmentation performance of the model. With one-domain source site medical images labeled, scans from other medical institutions of different renal artery periods can obtain competitive segmentation results close to the fully-supervised approach, which has great clinical significance.

In clinical practice, as a novel O2M-UDA framework, the proposed DyDA uses a dynamic strategy to minimize the distribution differences and gradually improve the model's adaptation ability, avoiding the additional annotation cost and training time cost. The proposed DyDA expands the application range of the model, breaking the privacy restrictions on medical data, which has important clinical significance. Additionally, the proposed DyDA does not ignore the performance on the source site making the model more practical.

The proposed innovations utilize the characteristics of the data distribution to iteratively expand the generalization boundary of the model and gradually adapt the model. The gradual expansion of the generalization boundary ensures the segmentation per-

formance of the model on the source site data and improves the adaptation ability of the model on the multiple-domain target site.

In addition, the proposed DyDA has a wide range of applications and is effective at minimizing distribution differences because we minimize the distribution differences by focusing on uncertainty reduction at two different levels, both in the detail and the entire level.

Compared with existing studies [35, 39, 40, 47, 49, 59], this study has the following advantages: **1)** We first explore the challenges existing in one-to-multiple medical image segmentation and proposes a novel framework for these challenges. **2)** DyDA proposes a new method of using data during training that dynamically uses data to gradually expand the generalization boundary of the model, thus improving the adaptation ability.

In future work, the division of the credible set based on the weighted credibility entropy evaluation has great potential to guide the segmentation performance on the target site of the segmentation model. Therefore, it is meaningful to explore more topics such as how to search hyperparameter combinations quickly, how to design a better credibility evaluation method based on specific tasks, and how to design a more generalized dynamic strategy to further improve the segmentation performance on the target site.

Acknowledgments

We thank the Big Data Computing Center of Southeast University for providing the facility support on the numerical calculations in this paper.

References

[1] Z.-H. Zhou, Open-environment machine learning, National Science Review 9 (8), nwac123. arXiv:https://academic.oup.com/nsr/article-pdf/9/8/nwac123/45472108/nwac123.pdf, doi:10.1093/nsr/nwac123.
URL https://doi.org/10.1093/nsr/nwac123

[2] J. Quinonero-Candela, M. Sugiyama, A. Schwaighofer, N. D. Lawrence, Dataset shift in machine learning, Mit Press, 2008.

[3] C. S. Perone, P. Ballester, R. C. Barros, J. Cohen-Adad, Unsupervised domain adaptation for medical imaging segmentation with self-ensembling, NeuroImage 194 (2019) 1–11.

[4] F. Prados, J. Ashburner, C. Blaiotta, T. Brosch, J. Carballido-Gamio, M. J. Cardoso, B. N. Conrad, E. Datta, G. Dávid, B. De Leener, et al., Spinal cord grey matter segmentation challenge, Neuroimage 152 (2017) 312–329.

[5] H. Guan, M. Liu, Domain adaptation for medical image analysis: a survey, IEEE Transactions on Biomedical Engineering 69 (3) (2021) 1173–1185.

[6] M. Kim, J. Zuallaert, W. De Neve, Towards novel methods for effective transfer learning and unsupervised deep learning for medical image analysis, in: Doctoral Consortium (DCBIOSTEC 2017), 2017, pp. 32–39.

[7] J. Yang, N. C. Dvornek, F. Zhang, J. Chapiro, M. Lin, J. S. Duncan, Unsupervised domain adaptation via disentangled representations: Application to cross-modality liver segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2019, pp. 255–263.

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

[9] Y. Tang, Y. Tang, V. Sandfort, J. Xiao, R. M. Summers, Tuna-net: Task-oriented unsupervised adversarial network for disease recognition in cross-domain chest x-rays, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2019, pp. 431–440.

[10] F. Pan, I. Shin, F. Rameau, S. Lee, I. S. Kweon, Unsupervised intra-domain adaptation for semantic segmentation through self-supervision, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3764–3773.

[11] D. Guan, J. Huang, S. Lu, A. Xiao, Scale variance minimization for unsupervised domain adaptation in image segmentation, Pattern Recognition 112 (2021) 107764. doi:https://doi.org/10.1016/j.patcog.2020.107764.
URL https://www.sciencedirect.com/science/article/pii/S0031320320305677

[12] L. Cui, X. Chen, D. Zhang, S. Liu, M. Li, M. Zhou, Multi-domain adaptation for smt using multi-task learning, in: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 2013, pp. 1055–1065.

[13] C. E. Shannon, A mathematical theory of communication, The Bell system technical journal 27 (3) (1948) 379–423.

[14] C. Bian, C. Yuan, J. Wang, M. Li, X. Yang, S. Yu, K. Ma, J. Yuan, Y. Zheng, Uncertainty-aware domain alignment for anatomical structure segmentation, Medical Image Analysis 64 (2020) 101732.

[15] T.-H. Vu, H. Jain, M. Bucher, M. Cord, P. Pérez, Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 2517–2526.

[16] L. Yu, S. Wang, X. Li, C.-W. Fu, P.-A. Heng, Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2019, pp. 605–613.

[17] J. T. Springenberg, Unsupervised and semi-supervised learning with categorical generative adversarial networks, arXiv preprint arXiv:1511.06390.

[18] T. Nair, D. Precup, D. L. Arnold, T. Arbel, Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation, Medical image analysis 59 (2020) 101557.

[19] I. Białynicki-Birula, J. Mycielski, Uncertainty relations for information entropy in wave mechanics, Communications in Mathematical Physics 44 (2) (1975) 129–132.

[20] J. Huang, D. Guan, A. Xiao, S. Lu, Rda: Robust domain adaptation via fourier adversarial attacking, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 8988–8999.

[21] T. Dietterich, Overfitting and undercomputing in machine learning, ACM computing surveys (CSUR) 27 (3) (1995) 326–327.

[22] T. Heimann, P. Mountney, M. John, R. Ionasec, Learning without labeling: Domain adaptation for ultrasound transducer localization, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2013, pp. 49–56.

[23] R. Bermúdez-Chacón, C. Becker, M. Salzmann, P. Fua, Scalable unsupervised domain adaptation for electron microscopy, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2016, pp. 326–334.

[24] P. Isola, J.-Y. Zhu, T. Zhou, A. A. Efros, Image-to-image translation with conditional adversarial networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1125–1134.

[25] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. Efros, T. Darrell, Cycada: Cycle-consistent adversarial domain adaptation, in: International conference on machine learning, PMLR, 2018, pp. 1989–1998.

[26] W. Hong, Z. Wang, M. Yang, J. Yuan, Conditional generative adversarial network for structured domain adaptation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1335–1344.

[27] M. Chen, H. Xue, D. Cai, Domain adaptation for semantic segmentation with maximum squares loss, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 2090–2099.

[28] Y. Zou, Z. Yu, B. Kumar, J. Wang, Unsupervised domain adaptation for semantic segmentation via class-balanced self-training, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 289–305.

[29] S. Dai, K. Sohn, Y.-H. Tsai, L. Carin, M. Chandraker, Adaptation across extreme variations using unlabeled domain bridges, arXiv preprint arXiv:1906.02238.

[30] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, Advances in neural information processing systems 27.

[31] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, D. Krishnan, Unsupervised pixel-level domain adaptation with generative adversarial networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 3722–3731.

[32] J.-Y. Zhu, T. Park, P. Isola, A. A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2223–2232.

[33] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, V. Lempitsky, Domain-adversarial training of neural networks, The journal of machine learning research 17 (1) (2016) 2096–2030.

[34] E. Tzeng, J. Hoffman, K. Saenko, T. Darrell, Adversarial discriminative domain adaptation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017, pp. 7167–7176.

[35] S. Fu, J. Chen, L. Lei, Cooperative attention generative adversarial network for unsupervised domain adaptation, Knowledge-Based Systems (2022) 110196doi:https://doi.org/10.1016/j.knosys.2022.110196.
URL https://www.sciencedirect.com/science/article/pii/S0950705122012928

[36] S. Sankaranarayanan, Y. Balaji, A. Jain, S. N. Lim, R. Chellappa, Learning from synthetic data: Addressing domain shift for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3752–3761.

[37] Y.-H. Tsai, W.-C. Hung, S. Schulter, K. Sohn, M.-H. Yang, M. Chandraker, Learning to adapt structured output space for semantic segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7472–7481.

[38] J.-C. Su, Y.-H. Tsai, K. Sohn, B. Liu, S. Maji, M. Chandraker, Active adversarial domain adaptation, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2020, pp. 739–748.

[39] Y. Wen, L. Chen, Y. Deng, Z. Zhang, C. Zhou, Pixel-wise triplet learning for enhancing boundary discrimination in medical image segmentation, Knowledge-Based Systems 243 (2022) 108424. doi:https://doi.org/10.1016/j.knosys.2022.108424.
URL https://www.sciencedirect.com/science/article/pii/S0950705122001708

[40] S. Aja-Fernández, A. H. Curiale, G. Vegas-Sánchez-Ferrero, A local fuzzy thresholding methodology for multiregion image segmentation, Knowledge-Based Systems 83 (2015) 1–12. doi:https://doi.org/10.1016/j.knosys.

2015.02.029.
URL https://www.sciencedirect.com/science/article/pii/S095070511500129X

[41] B. N. Li, C. K. Chui, S. Chang, S. H. Ong, Integrating spatial fuzzy clustering with level set methods for automated medical image segmentation, Computers in biology and medicine 41 (1) (2011) 1–10.

[42] D. D. Patil, S. G. Deore, Medical image segmentation: a review, International Journal of Computer Science and Mobile Computing 2 (1) (2013) 22–27.

[43] D. L. Pham, C. Xu, J. L. Prince, A survey of current methods in medical image segmentation, Annual review of biomedical engineering 2 (3) (2000) 315–337.

[44] J. Rogowska, Overview and fundamentals of medical image segmentation, Handbook of medical imaging, processing and analysis (2000) 69–85.

[45] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer, 2015, pp. 234–241.

[46] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, K. H. Maier-Hein, nnu-net: a self-configuring method for deep learning-based biomedical image segmentation, Nature Methods (2020) 1–9.

[47] Z. Han, M. Jian, G.-G. Wang, Convunext: An efficient convolution neural network for medical image segmentation, Knowledge-Based Systems 253 (2022) 109512.

[48] Y. He, G. Yang, J. Yang, R. Ge, Y. Kong, X. Zhu, S. Zhang, P. Shao, H. Shu, J.-L. Dillenseger, et al., Meta grayscale adaptive network for 3d integrated renal structures segmentation, Medical Image Analysis 71 (2021) 102055.

[49] Y. Yang, T. Yan, X. Jiang, R. Xie, C. Li, T. Zhou, Mh-net: Model-data-driven hybrid-fusion network for medical image segmentation, Knowledge-Based Systems 248 (2022) 108795.

[50] Y. He, G. Yang, J. Yang, Y. Chen, Y. Kong, J. Wu, L. Tang, X. Zhu, J.-L. Dillenseger, P. Shao, S. Zhang, H. Shu, J.-L. Coatrieux, S. Li, Dense biased networks with deep priori anatomy and hard region adaptation: Semi-supervised learning for fine renal artery segmentation., Medical Image Analysis 63 (2020) 101722.

[51] Z. Dong, Y. He, X. Qi, Y. Chen, H. Shu, J.-L. Coatrieux, G. Yang, S. Li, Mnet: Rethinking 2d/3d networks for anisotropic medical image segmentation, Vienna, Austria, 2022, pp. 870 – 876.

[52] N. Heller, F. Isensee, K. H. Maier-Hein, X. Hou, C. Xie, F. Li, Y. Nan, G. Mu, Z. Lin, M. Han, et al., The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge, Medical Image Analysis 67 (2020) 101821.

[53] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P.-A. Heng, I. Cetin, K. Lekadir, O. Camara, M. A. G. Ballester, et al., Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved?, IEEE transactions on medical imaging 37 (11) (2018) 2514–2525.

[54] X. Zhuang, L. Li, Myocardial Pathology Segmentation Combining Multi-Sequence Cardiac Magnetic Resonance Images: First Challenge, MyoPS 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Proceedings, Vol. 12554, Springer Nature, 2020.

[55] A. Lalande, Z. Chen, T. Decourselle, A. Qayyum, T. Pommier, L. Lorgis, E. de la Rosa, A. Cochet, Y. Cottin, D. Ginhac, et al., Emidec: a database usable for the automatic evaluation of myocardial infarction from delayed-enhancement cardiac mri, Data 5 (4) (2020) 89.

[56] X. Zhu, H. Zhou, C. Yang, J. Shi, D. Lin, Penalizing top performers: Conservative loss for semantic segmentation adaptation, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 568–583.

[57] H. Abdi, L. J. Williams, Principal component analysis, Wiley interdisciplinary reviews: computational statistics 2 (4) (2010) 433–459.

[58] K. Krishna, M. Narasimha Murty, Genetic k-means algorithm, IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 29 (3) (1999) 433–439. doi:10.1109/3477.764879.

[59] Y. Liu, H. Wang, Z. Chen, K. Huangliang, H. Zhang, Transunet: Redesigning the skip connection to enhance features in medical image segmentation, Knowledge-Based Systems 256 (2022) 109859. doi:https://doi.org/10.1016/j.knosys.2022.109859.
URL https://www.sciencedirect.com/science/article/pii/S0950705122009522

CRediT authorship contribution statement

Ziyue Jiang: Conceptualization, Methodology, Software, Writing - Review & Editing.
Yuting He: Reviewing & Editing, Formal analysis.
Shuai Ye: Investigation, Software, Resources.
Peifeng Shao: Visualization, Resources.
Xiaomei Zhu: Data Curation, Visualization.
Yi Xu: Data Curation, Supervision.
Yang Chen: Investigation, Supervision.
Jean-Louis Coatrieux: Supervision.
Shuo Li: Review & Editing, Project administration.
Guanyu Yang: Project administration, Conceptualization, Review & Editing, Funding acquisition.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: