Medical Image Analysis 82 (2022) 102617

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Source-free domain adaptation for image segmentation

Mathilde Bateson [a,*], Hoel Kervadec [a,b], Jose Dolz [a,b], Hervé Lombaert [a], Ismail Ben Ayed [a,b]

[a] ÉTS Montréal, 1100 Notre-Dame St W, Montreal, Quebec H3C 1K3, Canada
[b] CRCHUM, 900 R. Saint-Denis, Montréal, QC H2X 0A9, Canada

ARTICLE INFO

Keywords:
Segmentation
Deep networks
Source-free domain adaptation
Shannon entropy
Mutual information
Prior knowledge

ABSTRACT

Domain adaptation (DA) has drawn high interest for its capacity to adapt a model trained on labeled source data to perform well on unlabeled or weakly labeled target data from a different domain. Most common DA techniques require concurrent access to the input images of both the source and target domains. However, in practice, privacy concerns often impede the availability of source images in the adaptation phase. This is a very frequent DA scenario in medical imaging, where, for instance, the source and target images could come from different clinical sites. We introduce a source-free domain adaptation for image segmentation. Our formulation is based on minimizing a label-free entropy loss defined over target-domain data, which we further guide with weak labels of the target samples and a domain-invariant prior on the segmentation regions. Many priors can be derived from anatomical information. Here, a class-ratio prior is estimated from anatomical knowledge and integrated in the form of a Kullback–Leibler (KL) divergence in our overall loss function. Furthermore, we motivate our overall loss with an interesting link to maximizing the mutual information between the target images and their label predictions. We show the effectiveness of our prior-aware entropy minimization in a variety of domain-adaptation scenarios, with different modalities and applications, including spine, prostate and cardiac segmentation. Our method yields comparable results to several state-of-the-art adaptation techniques, despite having access to much less information, as the source images are entirely absent in our adaptation phase. Our straightforward adaptation strategy uses only one network, contrary to popular adversarial techniques, which are not applicable to a source-free DA setting. Our framework can be readily used in a breadth of segmentation problems, and our code is publicly available: https://github.com/mathilde-b/SFDA.

1. Introduction

1.1. Motivation

Unprecedented advances in visual recognition tasks have been possible thanks to the improvements in hardware, novel deep architectures and availability of large annotated datasets. Deep Convolutional Neural Networks (CNNs) can provide powerful image representations when trained on huge amounts of labeled images, which can be used in a breadth of computer vision problems. For instance, CNNs have outstandingly improved automated methods for segmentation in many natural and medical imaging problems (Litjens et al., 2017). A major impediment of such supervised models is that they require large amounts of training data built with scarce expert knowledge and labor-intensive, pixel-level annotations. Typically, segmentation ground truth is available for limited data, and supervised models are seriously challenged with new samples (target data) that differ from the labeled training samples (source data). In medical imaging, for instance, the data distribution may vary significantly across different vendors, machines, image modalities and acquisition protocols, as illustrated on Fig. 1. Such domain shifts between different scans introduce a significant variability in the appearances of the target regions, impeding the generalization of CNN segmentation models. There has been an ongoing research effort towards improving the performance of models across domains, without retraining them nor labeling entire datasets in new target domains, which would be impractical in medical imaging (Cheplygina et al., 2019).

Domain Adaptation (DA) addresses the transferability of a model trained on an annotated source domain to another target domain with no, or minimal annotations. With the advent of Generative Adversarial Networks (GANs) (Goodfellow et al., 2014), adversarial-learning techniques widely dominate the recent literature in domain adaptation for segmentation. One major limitation of adversarial techniques is that, by design, they require concurrent access to both the source and target data during the adaptation phase. More generally, other

* Corresponding author.
E-mail address: mathilde.bateson.1@ens.etsmtl.ca (M. Bateson).

https://doi.org/10.1016/j.media.2022.102617
Received 21 July 2021; Received in revised form 25 July 2022; Accepted 2 September 2022
Available online 16 September 2022

recent approaches to DA, such as those based on self-training, also use both source and target data during adaptation. However, in many medical imaging scenarios, the source data may not be available in the adaptation phase. This involves, for example, confidentiality reasons, loss or corruption of the source data, or computational constraints for real-time applications.

Instead, we tackle *Source-Free Domain Adaptation*, where the source data is not accessible during the adaptation phase. Our adaptation relies on minimizing a loss containing the Shannon entropy of predictions and a class-ratio prior on the target domain (i.e., the proportion of a region in an entire image). This loss implicitly matches the prediction statistics of the source and target domains, thereby removing the need for complex two-step adversarial training as in GANs. Moreover, we show the robustness of our framework to substantial uncertainty in the class-ratio prior, and give an information-theoretic perspective of our loss. Our method enables to embed approximate anatomical knowledge, and to leverage weak labels of the target samples in the form of image-level tags for segmentation tasks.

1.2. Related work

Among the earliest works aiming to address domain-shift problems, Crammer et al. (2007), Ben-David et al. (2010), Pan and Yang (2010) propose to find a mapping of data distributions from a source to a target. More precisely, to tackle the discrepancy between the two domains, the learning process exploits the differences of data distributions across domains, yielding domain-invariant features. The main idea is to find an intermediate feature space where the marginal distribution of the source is similar to the target. Thus, we can assume that, in this intermediate representation, the prediction function is the same across source and target domains. This results in models that can be trained using annotated datasets from the source domain along with unlabeled or weakly labeled target data, with a strong cross-domain generalization ability.

Adversarial methods: Inspired by this assumption, recent works have focused on leveraging deep learning models to extract domain invariant features from input images (Ganin and Lempitsky, 2015; Long et al., 2015; Tzeng et al., 2015). Particularly, most of the existing research exploits deep adversarial training (Ganin et al., 2016) in a wide range of applications and problems, such as classification (Tzeng et al., 2017; Wachinger and Reuter, 2016; Tulder and Bruijne, 2016; Sankaranarayanan et al., 2018) or segmentation (Kamnitsas et al., 2017; Hoffman et al., 2018; Huo et al., 2018; Javanmardi and Tasdizen, 2018; Tsai et al., 2018; Zhang et al., 2018a; Zhao et al., 2018). These methods either follow a generative approach, by transforming images from one domain to the other (Zhu et al., 2017; Huo et al., 2019), or minimize the discrepancy in the feature or output spaces learnt by the model (Dou et al., 2019; Tzeng et al., 2017; Tsai et al., 2018). As these two perspectives are in essence complementary, the recent methods achieve state-of-the-art performances for adapting semantic segmentation in natural (Hoffman et al., 2018; Zhang et al., 2018b) and medical images (Chen et al., 2020) by combining image- and feature-alignment strategies. One major limitation of adversarial techniques is that, by design, they require concurrent access to both the source and target data during the adaptation phase.

Self-training: Amongst alternative approaches to adversarial techniques, self-training (Zou et al., 2018) and the closely-related entropy minimization (Vu et al., 2019; Wu et al., 2020; Morerio et al., 2018) were investigated in computer vision. As confirmed by the low entropy prediction maps in Fig. 1, a model trained on an imaging modality tends to produce very confident predictions on within-sample examples, whereas uncertainty remains high on unseen modalities. Moreover, the entropy maps can identify inaccurate segmentation regions in these target examples. As a result, enforcing a higher confidence of predictions in the target domain would help decreasing this performance gap. This is the underlying motivation for entropy minimization, which was first introduced in the contexts of semi-supervised (Grandvalet and Bengio, 2004) and unsupervised (Krause et al., 2010) learning. To prevent the well-known collapse of entropy minimization to a trivial solution with a single class, the recent domain-adaptation methods in Vu et al. (2019), Wu et al. (2020) further incorporate a criterion encouraging diversity in the prediction distributions, while (Bian et al., 2020) minimize the uncertainty measured as the variance of the network's output, in combination with adversarial learning. However, similarly to adversarial approaches, all these uncertainty-based methods require access to the source data, both the images and labels, during the adaptation phase. The source data is used to compute the standard supervised cross-entropy loss and/or used in an adversarial adaptation, to prevent trivial solutions that are obtained by minimizing uncertainty on the unlabeled target images.

Test-time Adaptation: Closest to our work, test-time domain adaptation (TTA) was introduced to improve generalization to new and different data, possibly a single data point, at test times. Most TTA methods comply with the SFDA setting: they relieve the need for accessing source domain data after the source training phase. Initial SFDA attempts addressed adapting classification tasks (Nath Kundu et al., 2020; Liang et al., 2020), either by using generative image translation (Benaim and Wolf, 2018) or self-supervision (Sun et al., 2020; Wang et al., 2021). Extensions to segmentation problems (Karani et al., 2021; He et al., 2020, 2021) alter the source-domain training with auxiliary branches used to align the target and source domains in the pixel, network-feature, and/or network-output spaces. A drawback of these methods is that the source training phase is non-standard (ex. require training an additional denoising network, Karani et al., 2021) and involve complex training and/or adaptation schemes. Varsavsky et al. (2020) proposed a test-time adaptation based on domain adversarial learning, which is adapted to a single target-domain subject, but is not source-free.

Domain Randomization Recent work Billot et al. (2020, 2021) has investigated the possibility to segment scans of arbitrary contrasts and resolutions by training with synthetic intensity images. These methods also comply with the source-free domain adaptation scenario.

Weakly supervised segmentation in medical imaging: To alleviate the burden of pixel-wise annotation, weakly supervised learning has become a popular strategy. In this setting, the supervision received by the segmentation network may come in the form of image-level tags (Wu et al., 2019; Ouyang et al., 2019; Patel and Dolz, 2022), bounding boxes (Rajchl et al., 2016; Kervadec et al., 2020), points (Khan et al., 2019; Dorent et al., 2021), scribbles (Tang et al., 2018), target size (Jia et al., 2017; Kervadec et al., 2019b) or, more recently, shape descriptors (Kervadec et al., 2021). On the one hand, approaches that rely on image-level tags typically use class-activation maps (Selvaraju et al., 2017), which are deployed to generate pseudo-labels, mimicking fully-supervised learning. On the other hand, knowledge-driven approaches typically embed prior-knowledge, such as the target size or location, in the learning objective. Furthermore, while most prior literature relies on in-distribution data, a very few attempts investigated domain adaptation in a weakly-supervised setting (Cheplygina et al., 2019; Bateson et al., 2021; Paul et al., 2020; Dorent et al., 2020). These works have shown promising results, especially when dealing with scarce data or severe domain shifts.

Leveraging the target class-ratio as a prior has shown a great potential to guide the training of segmentation models when dealing with limited supervision, including weakly (Jia et al., 2017; Kervadec et al., 2019b), semi-supervised (Zhou et al., 2019; Kervadec et al., 2019a) or few-shot (Boudiaf et al., 2021) learning. In the presence of domain shifts, several recent works have also resorted to this prior as a source of additional supervision (Vu et al., 2019; Zhang et al., 2019; Bateson et al., 2021). An important difference, however, is that prior works require accessing the source data. Indeed, their learning objectives include a cross-entropy loss over the labeled source images during the training of the adaptation phase. This contrasts with our setting, as we relax this requirement.

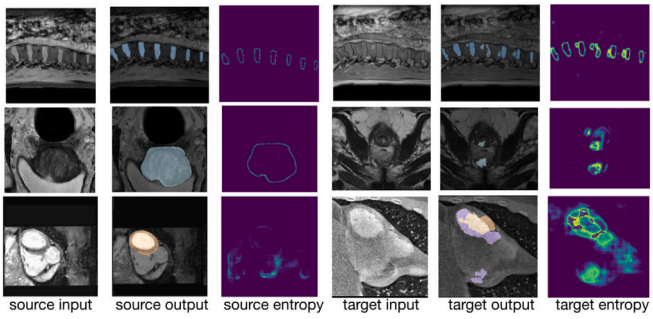

Fig. 1. Visualization of severe domain shifts between source and target modalities along with their corresponding predicted segmentation and entropy maps in three applications. Top: 2 spine images from Water (left) and In-Phase (right) MRI, with the intervertebral disks depicted in blue and the background in black. Middle: 2 prostate MRI images from different sites. Bottom: 2 cardiac images from MRI (left) and CT (right). The cardiac structures of AA, LV and MYO are depicted in blue, purple and brown, respectively. The domain shift in the target causes a drop in confidence and accuracy.

1.3. Contributions

We propose a *Source-Free Domain Adaptation* formulation (SFDA) tailored to a setting where the source data is unavailable, neither its images nor its labeled masks, during the training of the adaptation phase. Instead, our method only requires the parameters of a model previously trained on the source data as an initialization; moreover, it does not use auxiliary branches or additional tasks trained on the source domain, contrary to previous SFDA methods (Karani et al., 2021; He et al., 2021, 2020). Our formulation is based on a minimization of a label-free entropy loss defined over the target-domain data, which we further guide with a domain-invariant prior on the segmentation regions. To facilitate adaptation, we leverage weak supervision in the form of image-level tags in the target domain. Furthermore, we provide an interesting connection between our loss and the mutual information between the target images and their label predictions.

We report a comprehensive set of experiments and comparisons with state-of-the-art domain-adaptation methods, which shows the effectiveness of our prior-aware entropy minimization in three applications: the adaptation of spine segmentation across different MRI modalities, the adaptation of prostate segmentation in MRI modalities across different sites and machines, and the adaptation of cardiac segmentation from MRI to CT. Surprisingly, even though our method does not have access to the source data during adaptation, it achieves comparable or even better performances than several state-of-the-art methods (Zhang et al., 2019; Tsai et al., 2018; Tzeng et al., 2017; Ganin et al., 2016; Zhu et al., 2017; Dou et al., 2019; Luo et al., 2019; Chang et al., 2019; Li et al., 2019), while greatly improving the confidence of network predictions.

A preliminary conference version of this work has appeared at MICCAI 2020 (Bateson et al., 2020). This journal version provides (1) a new loss to tackle source-free adaptation, with an interesting mutual-information perspective and better gradient dynamics than the one introduced in Bateson et al. (2020); (2) two new applications; (3) ablation studies; and (4) the introduction of anatomical knowledge to estimate the class-ratio priors, which demonstrates the practical usefulness of our method and its robustness to uncertainty in estimating the priors. Specifically, unlike Bateson et al. (2020), we perform comprehensive evaluations in a setting where the class-ratio priors of the target regions are not estimated by an auxiliary network, but rather derived from textbook anatomical knowledge, even with substantial imprecision. We argue that such an approach offers a great potential in multiple clinical settings, particularly when access to source data is compromised. Our framework can be readily used for adapting a breadth of segmentation problems, with the code made publicly available.[1]

The contributions of this paper can be summarized as follows:

1. We tackle Source-Free Domain Adaptation (SFDA), a setting where the source data is unavailable, neither its images nor labeled masks, during the training of the adaptation phase. Our formulation allows SFDA with no modification to the source training.
2. We propose a novel loss defined over the unlabeled target-domain data, which integrates the Shannon entropy with a Kullback–Leibler divergence matching the class-ratios of the segmentation regions to an anatomical prior. Furthermore, we motivate our loss with an interesting link to maximizing the mutual information between the target images and their latent labels.
3. We extensively validate our method on three DA datasets. The results show that our framework can effectively and efficiently address the domain shift problem without accessing the source data during the adaptation phase.

2. Method

We consider a set $\mathcal{S}$ of source images $I_s : \Omega_s \subset \mathbb{R}^d \rightarrow \mathbb{R}$, $d \in \{2\}$, $s = 1, \dots, S$. The ground-truth K-class segmentation of I_s can be written, for each pixel (or voxel) $i \in \Omega_s$, as a simplex vector $\mathbf{y}_s(i) = (y_s^1(i), \dots, y_s^K(i)) \in \{0,1\}^K$. For domain adaptation (DA) problems, typically, a deep network is first trained on the source domain only, by minimizing a standard supervised loss with respect to network parameters θ:

$$\mathcal{L}_s\left(\theta, \Omega_s\right) = \frac{1}{S}\sum_{s=1}^{S}\frac{1}{|\Omega_s|}\sum_{i\in\Omega_S}\ell\left(\mathbf{y}_s(i), \mathbf{p}_s(i,\theta)\right) \tag{1}$$

where $\mathbf{p}_s(i,\theta) = (p_s^1(i,\theta), \dots, p_s^K(i,\theta)) \in [0,1]^K$ is the softmax output of the network at i in image I_s, and here we take ℓ as the standard cross-entropy loss: $\ell(\mathbf{y}_s(i), \mathbf{p}_s(i,\theta)) = -\sum_k y_s^k(i)\log p_s^k(i,\theta)$.

The adaptation phase is then initialized with the network parameters $\tilde{\theta}$ obtained from the source training phase. Given a set $\mathcal{T}$ of images in the target domain, $I_t : \Omega_t \subset \mathbb{R}^2 \rightarrow \mathbb{R}$, $t = 1, \dots, T$, the first loss term in our adaptation phase encourages high confidence in the softmax predictions of the target, which we denote $\mathbf{p}_t(i,\theta) = (p_t^1(i,\theta), \dots, p_t^K(i,\theta)) \in [0,1]^K$. This is done by minimizing a weighted Shannon entropy of each of these predictions:

$$\ell_{ent}(\mathbf{p}_t(i,\theta)) = -\nu_k\sum_k p_t^k(i,\theta)\log p_t^k(i,\theta) \tag{2}$$

where $\nu_k, k = 1, \dots, K$, are non-negative constants denoting class weights added to alleviate the burden of unbalanced class-ratios.

However, it is well-known from the semi-supervised and unsupervised learning literature (Grandvalet and Bengio, 2004; Krause et al., 2010; Jabi et al., 2021) that minimizing this entropy loss alone may result into trivial solutions, where the predictions are biased towards a single dominant class. To avoid such degenerate solutions, the recent domain-adaptation work of Vu et al. (2019), Wu et al. (2020) have integrated a standard supervised cross-entropy loss over the source data, such as in Eq. (1), when training during the adaptation phase. This, however, requires access to the source data, both its images and labels, during the adaptation phase. To remove this undesired requirement, we embed a domain-invariant prior knowledge to guide the unsupervised entropy training during the adaptation phase, which takes the form of a class-ratio prior (i.e., the proportion of a region in an entire image). The unknown true class-ratio prior for a class k and image I_t can be computed as follows: $\tau_{GT}(t,k) = \frac{1}{|\Omega_t|}\sum_{i\in\Omega_t} y_t^k(i)$.

[1] https://github.com/mathilde-b/SFDA

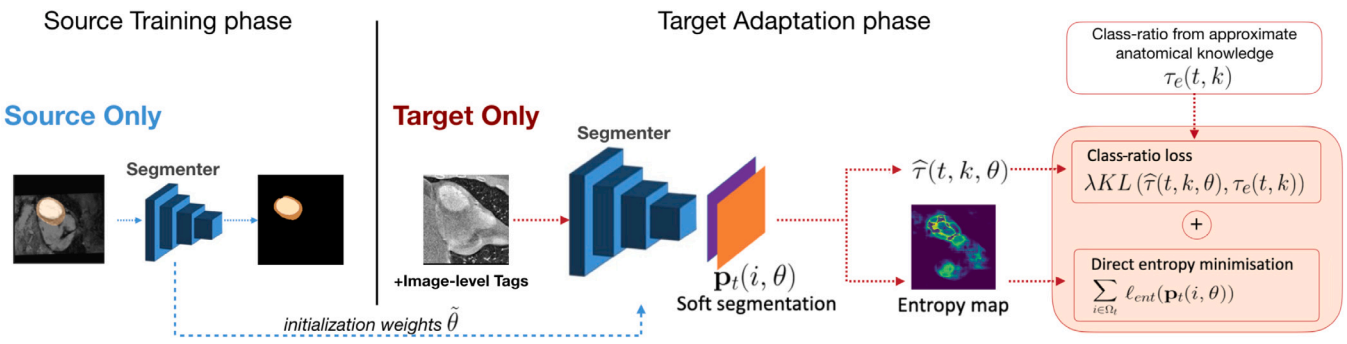

Fig. 2. Overview of our framework for Source-Free Domain Adaptation: we leverage entropy minimization and a class-ratio prior, to remove the need for a concurrent access to the source and target data.

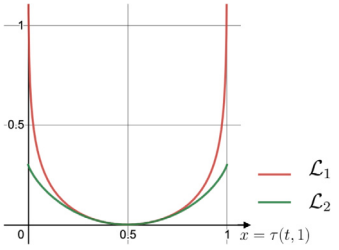

Fig. 3. Comparison of two class-prior losses in the scenario $K=2$, with the ground-truth class-ratio set to $\tau_{GT}(t,1)=0.5$: The plots illustrate better gradient dynamics of $\mathcal{L}_2$ at the vicinity of a class-ratio $\tau(t,1)=0$. Best seen in colors.

This gives the size of class k in image I_t over the image size. However, as the ground-truth labels are unavailable in the target domain, this prior cannot be computed directly. Instead, we estimate it with simple region statistics from anatomical prior knowledge, which we denote as $\tau_e(t,k)$. Furthermore, the class-ratio of the segmentation network output prediction can be computed as follows: $\widehat{\tau}(t,k,\theta)=\frac{1}{|\Omega_t|}\sum_{i\in\Omega_t}p_t^k(i,\theta)$.

We regularize the entropy in Eq. (2) with a Kullback–Leibler (KL) divergence matching these two class-ratios. Thus, our method minimizes the following overall loss during the training of the adaptation phase:

$$\min_{\theta}\sum_t\frac{1}{|\Omega_t|}\sum_{i\in\Omega_t}\ell_{ent}(\mathbf{p}_t(i,\theta))+\mathrm{KL}(\widehat{\tau}(t,\theta,\cdot),\tau_e(t,\cdot)) \tag{3}$$

where $\mathrm{KL}(\widehat{\tau}(t,\theta,\cdot),\tau_e(t,\cdot))=\widehat{\tau}(t,\theta,\cdot)\log\left(\frac{\widehat{\tau}(t,\theta,\cdot)}{\tau_e(t,\cdot)}\right)$.

Clearly, minimizing our overall loss in Eq. (3) during adaptation does not use the source images and labels. In the following, we discuss an interesting link between our loss in Eq. (3) and maximizing the mutual information between the target images and their network predictions. Fig. 2 shows the overview of the proposed framework.

2.1. Link to mutual-information maximization

Notice that the terms of the KL penalty in Eq. (3) are inverted compared to our initial formulation (*AdaEnt*), which we provided in the conference version of this work (Bateson et al., 2020); see Eq. (9). Besides the empirical motivation (as it will be shown in the experimental section hereafter), this is first and foremost motivated by theoretical results in information theory, as we link below Eq. (3) to maximizing the mutual information between the input images and their latent label predictions. The full proof is derived in Appendix B.

Let $\mathcal{I}(X,Y)$ denote the mutual information between two random variables X and Y:

$$\begin{aligned}\mathcal{I}(X;Y)&=H(Y)-H(Y\mid X)\\&=-\mathbb{E}_Y\left[\log\mathbb{E}_X[p(Y\mid X)]\right]+\mathbb{E}_{X,Y}[\log p(Y\mid X)]\end{aligned} \tag{4}$$

where $H(Y)$ is the entropy of Y, $H(Y\mid X)$ is the conditional entropy of Y given X, and $\mathbb{E}_X[p(Y\mid X)]$ is the marginal distribution of Y under the conditional model $p(Y\mid X)$.

We denote P_t the $K\times|\Omega_t|$ softmax prediction mask, i.e. matrix whose columns are the vectors of network outputs $\mathbf{p}_t(i,\theta), i\in\Omega_t$. Given the classical interpretation of the softmax predictions as probabilities: $p_t^k(i,\theta)=p(y_t^k(i)=1\mid I_t,\theta)$, the empirical class-ratio distribution is an estimate of the marginal distribution of P_t: $\hat{\tau}(t,\theta,\cdot)=\frac{1}{|\Omega_t|}\sum_{i\in\Omega_t}\mathbf{p}_t(i,\theta)=\mathbb{E}_{I_t}[p(P_t\mid I_t)]$. Therefore the empirical estimate of the mutual information between the images I_t and their softmax predictions, $P_t, t=1,\dots,T$, can be expressed as[2]:

$$\mathcal{I}_\theta=\frac{1}{T}\sum_t\underbrace{H\{\hat{\tau}(t,\theta,\cdot)\}}_{-\mathbb{E}_{P_t}\left[\log\mathbb{E}_{I_t}[p(P_t|I_t)]\right]}\underbrace{-\frac{1}{|\Omega_t|}\sum_{i\in\Omega_t}\ell_{ent}(\mathbf{p}_t(i,\theta))}_{-\mathbb{E}_{I_t,P_t}[\log p(P_t|I_t)]} \tag{5}$$

In the different context of discriminative clustering, Krause et al. (2010) draw a connection between maximizing the empirical estimate of the mutual information, as in Eq. (5), and a generalization of the mutual information based on the KL divergence, as in Eq. (3). Indeed, note that the following basic identity holds:

$$H\{\hat{\tau}(t,\theta,\cdot)\}\overset{c}{=}-KL\{\hat{\tau}(t,\theta,\cdot),U\} \tag{6}$$

where U is the uniform distribution over labels $\{1,\dots,K\}$. The term $KL\{\hat{\tau}(t,.,\theta),U\}$ is maximized when the class-ratio distribution is uniform. Instead, to integrate a prior about the class-ratio distribution, for each image I_t and class k, we can replace U by prior distribution $\tau_e(t,.)$ as follows:

$$\max_{\theta}\sum_t-KL\{\hat{\tau}(t,\theta,\cdot),\tau_e(t,\cdot)\}-\sum_t\frac{1}{|\Omega_t|}\sum_{i\in\Omega_t}\ell_{ent}(\mathbf{p}_t(i,\theta)) \tag{7}$$

which is equivalent to Eq. (3). Maximizing the mutual information between the images I_t and their softmax predictions $p_t(\theta)$ is a principled approach in unsupervised problems, such as unsupervised discriminative clustering (Krause et al., 2010; Jabi et al., 2021), further motivating our formulation, which we denote *AdaMI* in the following.

2.2. Choosing the penalty function

Given an image I_t, consider the penalty functions $\mathcal{L}_1$ (resp. $\mathcal{L}_2$) used in combination with entropy minimization in *AdaEnt* (resp. in *AdaMI*):

$$\mathcal{L}_1=\mathrm{KL}(\tau_e(t,\cdot),\widehat{\tau}(t,\theta,\cdot))$$
$$\mathcal{L}_2=\mathrm{KL}(\widehat{\tau}(t,\theta,\cdot),\tau_e(t,\cdot))$$

Fig. 3 shows the profile of these two regularizers as functions of the class-ratio for a binary-segmentation case, with a target foreground

[2] See details of proof in Appendix B

class-ratio set to 0.5. We see that $\mathcal{L}_2$ may be a better choice than $\mathcal{L}_1$ when the initial predictions of the network are extremely imbalanced. Indeed, note the gradient properties and stability at the vicinity of 0, i.e., when the predicted foreground class-ratio $\tau(t,1)$ is close to 0. We see that both first and second derivatives of the regularizer are unbounded for $\mathcal{L}_1$, but bounded and constant for $\mathcal{L}_2$. Our experiments confirm the superiority of the $\mathcal{L}_2$ regularizer, in terms of training stability and quantitative performance.

2.2.1. Estimating the class-ratio prior from anatomical knowledge

In Bateson et al. (2020), the ground-truth class-ratio is estimated through an auxiliary network trained with the source data. In a more general source-free scenario, only the weights $\tilde{\theta}$ of a network trained with the source data are available during the adaptation phase, and the class-ratio cannot be learnt, neither estimated from the source data. Therefore, we resort here to the more general case where the true class-ratio $\tau_{GT}(t,k)$ of each structure k in an image I_t is estimated from anatomical knowledge $\bar{\tau}_k$ available in the clinical literature (see A for our estimates from anatomical information).

For each 2D target image I_t and each structure k, the class-ratio used for adapting the segmentation network with Eq. (3) is obtained by adding weak supervision in the form of image-level tag information:

$$\tau_e(t,k) = \begin{cases} \bar{\tau}_k & \text{if region } k \text{ is within image } t. \\ 0 & \text{otherwise,} \end{cases} \tag{8}$$

Note that we use exactly the same class-ratio priors and weak supervision in our *AdaEnt* method, for a fair comparison.

3. Experiments and results

3.1. Experimental settings

3.1.1. Datasets

IVDM3Seg. The proposed SFDA method is first evaluated on the dataset from the MICCAI 2018 IVDM3Seg Challenge,[3] consisting of 16 3D multi-modality MRI datasets, collected from 8 subjects at two different stages to study inter-vertebral disk (IVD) degeneration. The scans were generated by a Dixon protocol with a 1.5 T Siemens MRI scanner, producing four aligned modalities. Scans are acquired in sagittal direction. Each volume has an anisotropic resolution of 2×1.25×1.25 mm/vx. The corresponding manual segmentations of the IVDs are also available. In our experiments, we set the water modality (Wat) as the source and the in-phase (IP) modality as the target domain. Therefore, in this setting, the source and target modalities are acquired from the same patient. From this dataset, 12 scans are used for training, one for validation, and the remaining 3 scans for testing. Images are normalized to zero mean and unit variance. Then, we performed a data augmentation based on affine transformations. The setting is binary segmentation (K=2).

NCI-ISBI13. We employ prostate T2-weighted MRIs from 2 different data sources with distribution shifts from the NCI-ISBI13 dataset, with their corresponding manual segmentations of the prostate region. The source dataset consists of 30 volumes from Radboud University Nijmegen Medical Centre, generated with a 3 T Siemens scanner. Each source volume has an anisotropic resolution of $0.4 \times 0.4 \times 3$ mm/vx. The target dataset consists of 30 volumes from Boston Medical Center generated with a 1.5 T Philips Achieva. Each target volume has an anisotropic resolution of 0.6-0.625 × 0.6-0.625 × 3.6-4 mm/vx. We use the publicly available pre-processed data provided by Liu et al. (2020), which resized each sample to 384 × 384 in axial plane, normalized it to zero mean and unit variance. We employed data augmentation based on affine transformations. We use 19 scans for training, one for validation, and the remaining 10 scans for testing.

[3] https://ivdm3seg.weebly.com/

MMWHS. We employ the 2017 Multi-Modality Whole Heart Segmentation (MMWHS) Challenge dataset for cardiac segmentation (Zhuang et al., 2019). The dataset consists of 20 MRI (source domain S) and 20 CT volumes (target domain T) of non-overlapping subjects, with their corresponding ground-truth masks. The source resolution is $0.78 \times 0.78 \times 1.6$ mm/vx, while the target resolution is around $1 \times 1 \times 1$ mm/vx. We adapt the segmentation network for parsing four cardiac structures: the Ascending Aorta (AA), the Left Atrium blood cavity (LA), the Left Ventricle blood cavity (LV) and the Myocardium of the left ventricle (MYO). We employ the pre-processed data provided by Dou et al. (2019), as well as their data split, with 14 subjects used for training, 2 for validation, and 4 for testing. All the data were normalized as zero mean and unit variance. In order to obtain a similar field of view for all volumes, they cropped the original scans to center the structures to segment using a 3D bounding box with a fixed coronal plane size of 256 × 256. Then, they performed a data augmentation based on affine transformations. We use this augmented dataset for our proposed method as well as the benchmark methods that we implemented.

3.1.2. Benchmark methods

The first experiment consists in evaluating the performance of the proposed approach on all three datasets against the following competing methods. Quantitative evaluations and comparisons with state-of-the-art methods are reported hereafter. We compare our proposed model *AdaMI* to the benchmark methods below, which have shown state-of-the-art performances for adapting segmentation networks.

Source-Free AdaEnt: We compare to the loss that we proposed in our original source-free domain adaptation (Bateson et al., 2020), denoted $AdaEnt$ in the following:

$$\sum_t \frac{1}{|\Omega_t|} \sum_{i \in \Omega_t} \ell_{ent}(\mathbf{p}_t(i,\theta)) + \lambda \mathrm{KL}(\tau_e(t,\cdot), \hat{\tau}(t,\theta,\cdot)) \tag{9}$$

Constrained Domain Adaptation: We compare to the method adopted in Bateson et al. (2021), referred to below as *CDA*:

$$\mathcal{L}_s(\theta,\Omega_s) + \frac{\lambda}{T} \sum_{t=1}^{T} [\tau_e(t,\cdot) - \hat{\tau}_t(t,\theta,\cdot)]^2$$

Curriculum Domain Adaptation: We denote *AdaSource* the method adopted in Zhang et al. (2019):

$$\mathcal{L}_s(\theta,\Omega_s) + \frac{\lambda}{T} \sum_{t=1}^{T} KL(\tau_e(t,\cdot), \hat{\tau}_t(t,\theta,\cdot))$$

Adversarial Domain Adaptation: We compare to *AdaptSegNet*, the method adopted in Tsai et al. (2018):

$$\mathcal{L}_s(\theta,\Omega_s) - \frac{\lambda}{T} \sum_{t=1}^{T} \sum_{i \in \Omega_T} \log\left(D\left(p_t(i,\theta)\right)^{(1)}\right)$$

where the adversarial loss maximizes the probability of a target sample being predicted as the source by a discriminator D.

Note that, for *CDA*, *AdaSource* and *AdaptSegNet*, the images from the source and target domains must be present concurrently during the adaptation phase. For *CDA* and *AdaSource*, the class-ratio is estimated through an auxiliary network trained with the source data and the weakly-supervised target data, as in Bateson et al. (2020).

We also compared to the following two source-free domain adaptation methods. The first is TTA (Karani et al., 2021), which trains an auxiliary denoising network on the source, then applies it to the noisy segmentations in the target. The second is Tent (Wang et al., 2021), which uses a simple entropy minimization, similarly to Eq. (2). Importantly, for both methods, instead of optimizing the whole segmentation network, only the normalization statistics and affine parameters of the network are updated, while the rest of the parameters are frozen.

A model trained on the source domain only using Eq. (1), *NoAdap*, is used as a lower bound. A model trained with the supervised cross-entropy loss on the target domain, referred to as *Oracle*, serves as an upper bound.

Finally, for the cardiac application, we also present benchmark results obtained in previous DA works (Bian et al., 2020; Dou et al., 2019), which we directly report in Table 2. The methods using AdaNet as the backbone were implemented in Dou et al. (2019), those with DeepLabV2 were implemented in Bian et al. (2020).

3.1.3. Evaluating robustness to class-ratio prior imprecision

In the following experiments, we investigate the impact on our SFDA approach of both precise and imprecise prior information about the class-ratios in the target domain. To this end, we train several models under the same setting, validating different values for the class-ratio priors on the target images. We illustrate on the challenging problem of segmenting cardiac structures, which have a high class-ratio variance amongst slices.

First, we investigate the capability of SFDA in the ideal setting when the precise size of the segmented region is known. To this end, for each image t and each structure k of the target domain, we use the following class-ratio derived from the ground-truth size:

$$\tau_{GT}(t,k) = \frac{1}{|\Omega_t|}\sum_{i\in\Omega_t} y_t^k(i) \tag{10}$$

This setting is hereafter referred as $AdaMI_{\tau_{GT}}$. This is followed by evaluating the robustness of our benchmarked method to a varying imprecision of the prior knowledge on the class-ratio prior, i.e., varying the size estimates of the segmented regions.For each image t and each structure k of the target domain (except the background), we use the following error on the class-ratio:

$$\tau(t,k) = (1 \pm \delta)\tau_e(t,k) \tag{11}$$

And then obtain the estimate the background estimation as: $\tau(t,0) = 1 - \sum_{k>1}\tau(t,k)$. We validate using imprecision errors varying with δ: $\{0.2, 0.4, 0.6\}$ and denote this setting $AdaMI_{\delta\tau}$ below.

3.1.4. Ablation study on target training dataset size

In this experiment, we study how much target training data is necessary for our method to achieve a successful adaptation. We train several models under the same setting, with a varying number of subjects in the target training dataset. This setting is hereafter referred as $AdaMI_{NT1}, AdaMI_{NT2}$..

3.1.5. Ablation study on the weak annotations in the target training dataset

Finally, we investigate the impact of removing the image-level tags in the target training dataset, i.e. a fully unsupervised source-free DA setting. Instead, we use an *estimation* of this tag derived from the network prediction, and select a *subset* of the target training images, while keeping the whole target validation and test set. More specifically, for each 2D target training image I_t and each structure k:

$$I_t \text{ is } \begin{cases} \text{selected, with } \tau_e(t,k) = \bar{\tau}_k & \text{if } \hat{\tau}(t,\tilde{\theta},k) > \frac{1}{4}\bar{\tau}_k. \\ \text{selected, with } \tau_e(t,k) = 0 & \text{if } \hat{\tau}(t,\tilde{\theta},k) = 0. \\ \text{discarded otherwise} \end{cases} \tag{12}$$

With $\tilde{\theta}$ the initial network parameters at the start of the adaptation phase. Note that the underlying motivation for this subset selection comes from the following observation: given a certain class label, the relative errors in size estimations for this class have a negative correlation with the true sizes. We then update this estimation once during training, at the epoch 100.

3.1.6. Training and implementation details

For all the methods, we employed UNet (Ronneberger et al., 2015), a widely used segmentation network due to its simplicity. The architecture used is the same one as for the original UNet paper. We use a 2D implementation for all applications. In the source training phase, a model is trained on the source data only with Eq. (1) for 150 epochs, a learning rate of 5×10^{-4}, and a learning rate decay of 0.9 every 20 epochs. The final model is used as initialization to the adaptation phase. In this phase, the model is adapted with Eq. (3), trained with the Adam optimizer (Kingma and Ba, 2014), for 150 epochs. For all applications, the initial learning rate is 1×10^{-6}, the weight decay is 10^{-3}, and the batch size is 24. The learning rate decay is 0.7 for the heart and prostate applications, and 0.2 for the spine one. It is applied every 20 epochs. For all methods, we pick the final model as the one achieving the best validation score. The weights from Eq. (2) are calculated as: $\nu_k = \frac{\bar{\tau}_k^{-1}}{\sum_k \bar{\tau}_k^{-1}}$.

3.1.7. Evaluation metrics

Our first evaluation metric is the Dice similarity coefficient (DSC), which measures the voxel-wise segmentation accuracy between the predicted and reference volumes. The second is the average symmetric surface distance (ASD), which calculates the average distances between the surface of the prediction mask and the ground truth. As the data is volumetric for all applications, these metrics are computed over the 3D segmentation masks.

3.2. Quantitative results

The quantitative performances of the different methods are presented in Table 1 for the spine and prostate images, and in Table 2 for the cardiac images.

No adaptation. First, we see that the models trained with full supervision on the source domain suffer from a drop in performance when used in a different target domain without any adaptation. In Fig. 4(c), it can be verified that the *NoAdap* is in an under-segmentation regime, with the predicted sizes of structures well below their true sizes. This validates that the predictions are biased towards the dominant class, which is the background here.

With adaptation. All models that use adaptation yield a substantial improvement over the lower baseline. For instance, on spine images, our model *AdaMI* reaches a Dice score (DSC) of 74.2%, representing 90% of the best-performing adaptation method, *AdapSegNet* (Tsai et al., 2018), which used the source data during adaptation. *AdaMI* yields a 1.17 ASD, which corresponds to an improvement by a multiplicative factor of 1.8 compared to the value for *NoAdap* (2.15 ASD). On prostate images, *AdaMI* reaches 79.5% DSC, 95% of the top performance *AdapSegNet*. An ASD of 3.92 is obtained, an improvement by a multiplicative factor of 3 compared to the value for *NoAdap* (10.59 ASD). Surprisingly, on cardiac images, where the domain shift is higher, *AdaMI* ranks second out of sixteen other adaptation techniques in terms of average DSC across cardiac structures, outperformed only by the recent method in Bian et al. (2020), a substantially more complex adaptation framework. Note that the quantitative results are not directly comparable between all models, since the backbone networks differ (see Table 2). These results show that having access to more information on source data does not necessarily help for the adaptation task. Finally, on all three applications, *AdaMI* outperforms the two other source-free domain adaptation methods. Specifically, *TTA* yields a smaller improvement than *AdaMI* on the spine and the prostate applications, and fails on the more difficult heart one. *Tent* only yields a small improvement in terms of Dice on all three applications.

AdaMI versus AdaEnt. The Dice scores (DSC) of our proposed *AdaMI* reach 85% of *Oracle*'s performance on spine images, 90% of its performance on prostate images, and 85% on cardiac images. This validates the efficiency of using a class-ratio prior matching with a KL divergence to prevent under-segmentation. Comparing *AdaMI* and *AdaEnt*, we see that on all three applications, *AdaMI* outperforms *AdaEnt* and shows better convergence properties (see Fig. 4(b)). Moreover, in Fig. 4(a), we can observe that *AdaEnt* reaches rapidly its highest validation DSC (first 20 epochs) before slowly decaying. Fig. 4(c) shows that the mean predicted size of structures jumps instantly from 50% below to 15%

Table 1
Performance comparison of the proposed formulation with different domain adaptation methods for spine (IVDM3Seg dataset, left) and prostate (NCI-ISBI13 dataset, right) segmentation, in terms of DSC (%) and ASD (vox).

Method	Source Free	Target Tags	Spine IVDs DSC	Spine IVDs ASD	Prostate DSC	Prostate ASD
NoAdap (lower bound)	✓	×	68.5	2.15	67.2	10.59
Oracle (upper bound)	✓	✓	87.5	0.38	88.4	1.81
AdaptSegNet Tsai et al. (2018)	×	×	**82.4**	**0.50**	**83.1**	**2.43**
AdaSource Zhang et al. (2019)	×	✓	75.9	0.99	76.3	3.93
CDA Bateson et al. (2021)	×	✓	75.7	0.86	77.9	3.28
TTA Karani et al. (2021)	✓	×	69.7	1.65	73.2	3.80
Tent Wang et al. (2021)	✓	×	68.8	1.84	68.7	5.87
Prior AdaEnt Bateson et al.	✓	✓	72.9	1.54	77.8	4.10
AdaMI (Ours)	✓	✓	74.2	1.17	79.5	3.92

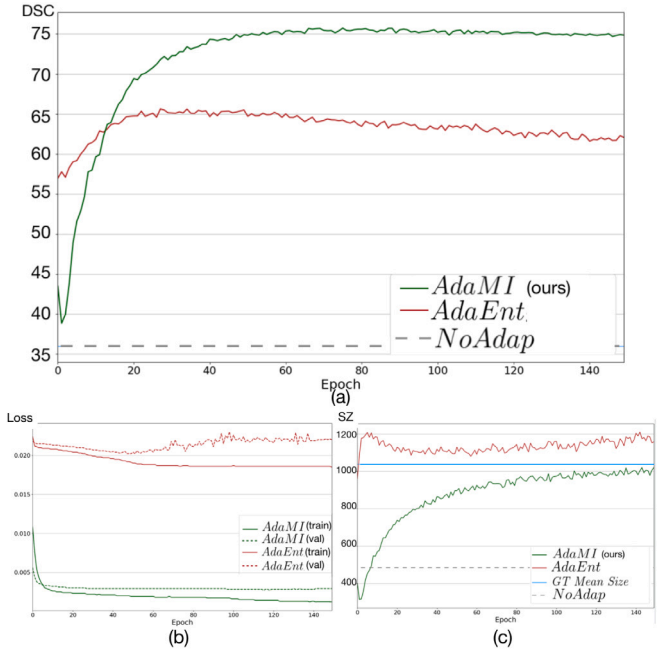

Fig. 4. Quantitative performance: (a) Evolution of DSC (%) and (b) Learning Curves and (c) mean ground truth sizes and predicted sizes (px) of cardiac structures segmentation masks over training epochs on target images from the validation set. Comparison of the proposed model *AdaMI*, and our previous *AdaEnt*.

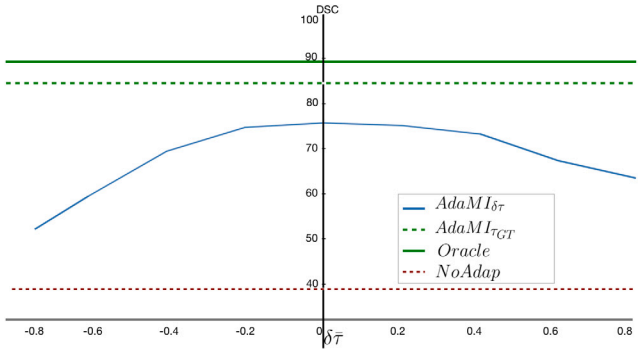

Fig. 5. Robustness performance: DSC (%) versus enforced relative size error in the class-ratio prior $\delta\tau$ for each structure for cardiac segmentation, showing robustness to imprecision in the prior. The DSC performance of the upper bounds *Oracle*, $AdaMI_{\tau_{GT}}$ and lower bound *NoAdap* are also indicated.

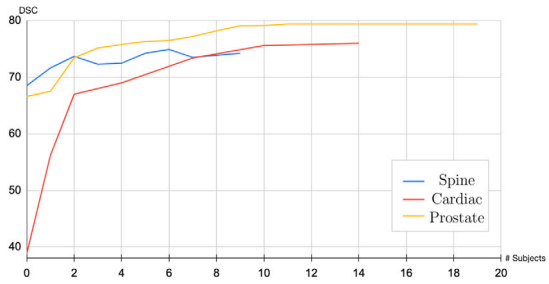

Fig. 6. Ablation performance: DSC (%) in target test set versus number of subject in the target training dataset for each application, showing the data efficiency of our method.

above the mean ground-truth sizes before stagnating. On the contrary, the performance of *AdaMI* improves steadily and the sizes of predicted structures grow progressively. This suggests that the inversion of the terms in the KL divergence in *AdaMI*, such as in Eq. (3), does help the learning process in domain adaptation, when compared to the original KL divergence in *AdaEnt* (see Section 2.2). Finally, the ASD values confirm the trend across the different models on cardiac images. Improvement over the lower baseline model (14.6 voxels) is substantial for *AdaEnt* (8.2 voxels), and even greater for *AdaMI* (5.6 voxels), with the greatest improvement occurring for AA and LA structures.

3.3. Ablation study on class-ratio precision

We also investigate the impact of imprecision in the target domain class-ratio prior on the quality of SFDA models. To this end, we validate a range of values in the estimations of class-ratios, as explained in Section 3.1.3. The results are reported for cardiac images in Fig. 5. First, in the ideal situation where the precise class-ratios are known, $AdaMI_{\tau_{GT}}$ reaches 84.5% DSC, representing 95% of the upper baseline, the *Oracle*. Then, we can see that our proposed method *AdaMI* is robust to large ranges of imprecision in class-ratio estimates. Indeed, a difference of ± 20% (resp. ± 40%) with our prior estimation in Section 2.2.1 only degrades the DSC by up to 1% (resp. 6%). Moreover, we see that an overestimation of the structure sizes leads to a better overall DSC than an underestimation, highlighting the well-known bias of Dice towards over-segmentation.

Finally, we emphasize that the class-ratio estimation used for a structure k is identical for all target images containing k. However, the true target class-ratios have high variance amongst slices. Thus the prior used in *AdaMI* is quite imprecise, which further confirms the robustness of our framework to class-ratio prior imprecision.

3.4. Ablation study on the size of the target training dataset

We also investigate how much weakly-labeled target training data is necessary for our SFDA model to achieve adaptation. To this end, we experiment with a varying number of subjects in the target training dataset. The results are reported in Fig. 6. We can see that our proposed method *AdaMI* is robust to large diminution of target dataset size.

Table 2
Performance comparison of the proposed formulation with different domain adaptation methods for cardiac segmentation, in terms of DSC (mean) and ASD (mean).

Methods	Source Free	Target Tags	Backbone	DSC (%)					ASD (vox)				
				AA	LA	LV	Myo	Mean	AA	LA	LV	Myo	Mean
NoAdap (lower bound)	✓	×	UNet	49.8	62.0	21.1	22.1	38.8	19.8	13.0	13.3	12.4	14.6
Oracle (upper bound)	✓	✓		91.9	88.3	91.0	85.8	89.2	3.1	3.4	3.6	2.2	3.0
AdaSource Zhang et al. (2019)	×	✓		79.0	77.9	64.4	61.3	70.7	6.5	7.6	7.2	9.1	7.6
CDA Bateson et al. (2021)	×	✓		77.3	72.8	73.7	61.9	71.4	**4.1**	6.3	6.6	**6.6**	5.9
TTA Karani et al. (2021)	✓	×		59.8	26.4	32.3	44.4	40.7	15.1	11.7	13.6	11.3	12.9
Tent Wang et al. (2021)	✓	×		55.4	33.4	63.0	41.1	48.2	18.0	8.7	8.1	10.1	11.2
Prior AdaEnt Bateson et al. (2020)	✓	✓		75.5	71.2	59.4	56.4	65.6	8.5	7.1	8.4	8.6	8.2
AdaMI (Ours)	✓	✓		**83.1**	**78.2**	**74.5**	**66.8**	**75.7**	5.6	**4.2**	**5.7**	6.9	**5.6**
AdaptSegNet Tsai et al. (2018)	×	×	DeepLabV2	65.4	80.6	81.4	69.3	74.2	8.1	5.3	4.0	3.6	5.2
BDL Li et al. (2019)	×	×		67.1	80.6	82.7	62.1	73.1	12.0	7.0	3.5	4.2	6.7
CLAN Luo et al. (2019)	×	×		63.8	79.9	**84.4**	66.8	73.7	9.1	5.3	**3.4**	**3.5**	5.3
DISE Chang et al. (2019)	×	×		71.8	82.2	83.7	60.8	74.6	6.7	4.7	3.8	7.7	5.7
SynSeg-Net Huo et al. (2019)	×	×		71.6	69.0	51.6	40.8	58.2	11.7	7.8	7.0	9.2	8.9
UADA Bian et al. (2020)	×	×		**84.1**	**88.3**	84.3	**71.4**	**82.1**	**3.9**	**3.5**	3.8	**3.7**	**3.7**
CyCADA Hoffman et al. (2018)	×	×	AdaNet	72.9	77.0	62.4	45.3	64.4	9.6	8.0	9.6	10.5	9.4
SIFA Chen et al. (2020)	×	×		**81.3**	**79.5**	**73.8**	**61.6**	**74.1**	**7.9**	**6.2**	**5.5**	**8.5**	**7.0**
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
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
DANN Ganin et al. (2016)	×	×		39.0	45.1	28.3	25.7	34.5	16.2	9.2	12.1	10.1	11.9
ADDA Tzeng et al. (2017)	×	×		47.6	60.9	11.2	29.2	37.2	13.8	10.2	NA	13.4	NA
Overall ranking of AdaMI (#/16)				2	7	6	3	**2**	3	2	7	6	**4**

Table 3
Performance of the proposed formulation obtained when removing the weak image-level annotations.

Method	Target tags	Dataset	DSC	ASD
AdaMI	✓	IVDM3Seg	74.2	1.17
		NCI-ISBI13	79.5	3.92
		MMWHS	75.7	5.6
$AdaMI_{unsupervised}$	×	IVDM3Seg	73.7	1.33
		NCI-ISBI13	71.8	7.49
		MMWHS	58.0	12.2

Indeed, with only 2 subjects, *AdaMI* is on par with most state-of-the art methods, reaching 67% DSC for the cardiac application, 74% DSC for the spine, and 73% for the prostate.

3.5. Ablation study on weak the annotations in the target training dataset

Finally, we investigate the more general scenario where images are fully unsupervised in the target domain. Particularly, we removed the target image tags for the adaptation phase as explained in Section 2.2.1. Results from this study are reported in Table 3. As expected, having image-level tag information helps all the models, which can be observed from the performance degradation compared to results in Tables 1 and 2. Indeed, the class-ratio estimation degrades without the image tag, and as a result, models using a class-ratio prior to guide adaptation also see their performance decrease. However, for the spine and the prostate application, the quantitative performance (73.7% DSC and 71.8% DSC respectively) remains well above the baseline, on par with most state-of-the art domain adaptation models. The removing of image-level Tags is more difficult for the heart application, as it is multi-class and has a big domain shift. However, results (58.0% DSC) stayed well above both the baseline and the two other SFDA methods, *Tent* and *TTA*.

3.6. Qualitative results

Qualitative segmentations and the corresponding entropy maps are shown for spine images in Fig. 7, for prostate images in Fig. 8, and for cardiac ones in Fig. 9. Without adaptation, the predictions of the network are either uncertain, as revealed by the high activation in the entropy maps of predictions (see top two lines in Fig. 9); or severely biased towards the dominant class, i.e. the background. This bias produces under-segmented or completely undetected structures (see the top four rows in Fig. 9). In all cases, the output segmentation masks are noisy, with very irregular edges. Benchmark adaptation models *CDA* and *AdaSource* are able to recover the structures in most examples. However, they display high uncertainty in the predictions, especially *CDA*. Interestingly, for some difficult slices, the segmentation results produced by our proposed SFDA model matches better with the ground-truth. For spine and prostate images, such examples are displayed in bottom two rows in Fig. 8. For cardiac images, the whole AA structure is better recovered (see middle two rows in Fig. 9), and the shapes and the boundary between the MYO and the LV structure are improved. Notably, in all applications, the entropy maps produced by *AdaMI* only show high activations along the borders of the predicted structures. These visual results further confirm the remarkable ability of *AdaMI* to produce accurate predictions with high confidence over existing approaches.

4. Discussion

We have introduced a source-free domain adaptation (SFDA) method to guide a segmentation network, trained on a source domain, to perform on a different target domain, without any access to the source-domain data in the adaptation phase. We have demonstrated the robustness of our SFDA approach on cross-modality spine MRI, cross-site prostate MRI, and MRI-to-CT cardiac adaptation.

Source-Free Domain Adaptation: Surprisingly, even though our model does not access the source data in the adaptation phase, it yields comparable or better performance than many state-of-the art adaptation approaches that do rely on the source data. It also outperforms two very recent source-free domain adaptation approaches, Wang et al. (2021), Karani et al. (2021). These works have stressed on the need for limited flexibility at test time, by freezing most parameters in the network, and adapting only the normalization and affine ones. Yet, in our three applications, we have found our proposed method, where the entire network is adapted, to be more efficient. Furthermore, our principled solution to source-free domain adaptation minimizes the uncertainty of the target domain predictions while preventing trivial solutions of single-class outputs via a KL regularizer that encourages target class-ratio (i.e region proportions). Using entropy minimization in combination with this regularizer, our formulation reaches 85%, 90% and 85% of full supervision in spine, prostate, and cardiac images

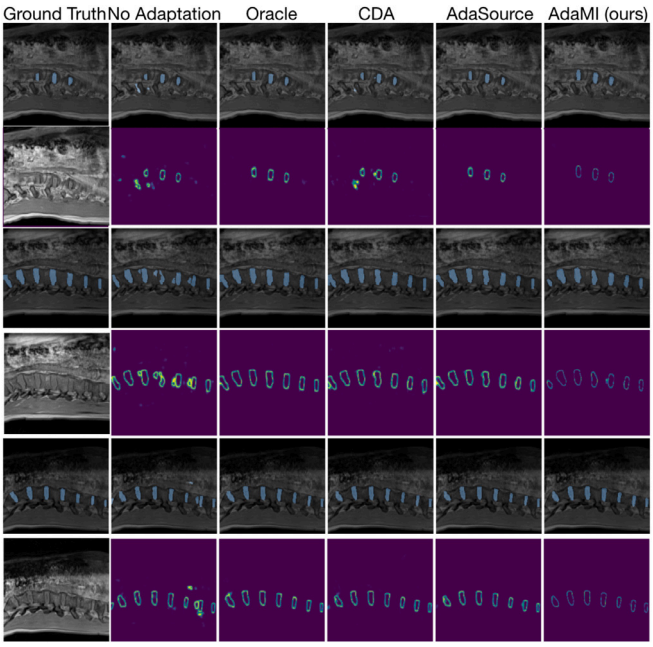

Fig. 7. Qualitative performance on spine MRI images: examples of the segmentations achieved by our formulation (*AdaMI*), benchmark models in Bateson et al. (2021), Zhang et al. (2019) and lower (*NoAdap*) and upper baselines (*Oracle*). First column shows an input slice and the corresponding semantic segmentation ground-truth. The other columns show segmentation results (top) along with prediction entropy maps produced by the different models (bottom).

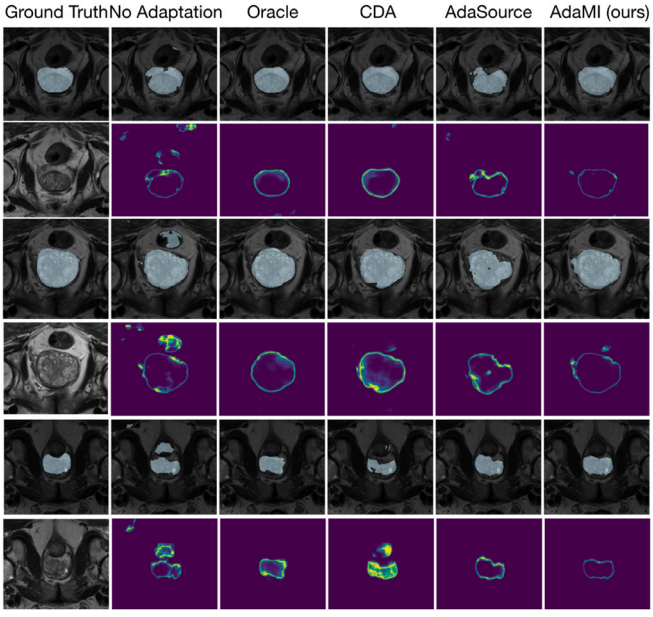

Fig. 8. Qualitative performance on prostate MRI images: examples of the segmentations achieved by our formulation (*AdaMI*), benchmark models in Bateson et al. (2021), Zhang et al. (2019) and lower (*NoAdap*) and upper baselines (*Oracle*).

respectively. Our qualitative results demonstrate the ability of SFDA to produce accurate predictions with high confidence.

Robustness: Our experiments have further confirmed the robustness of *AdaMI* to substantial prior imprecision, and that having a coarse knowledge of the target region proportions can be enough to guide adaptation. In our implementation, a class-ratio prior is derived from readily available anatomical reference values. This anatomical knowledge is combined with image-level tags to produce a very coarse

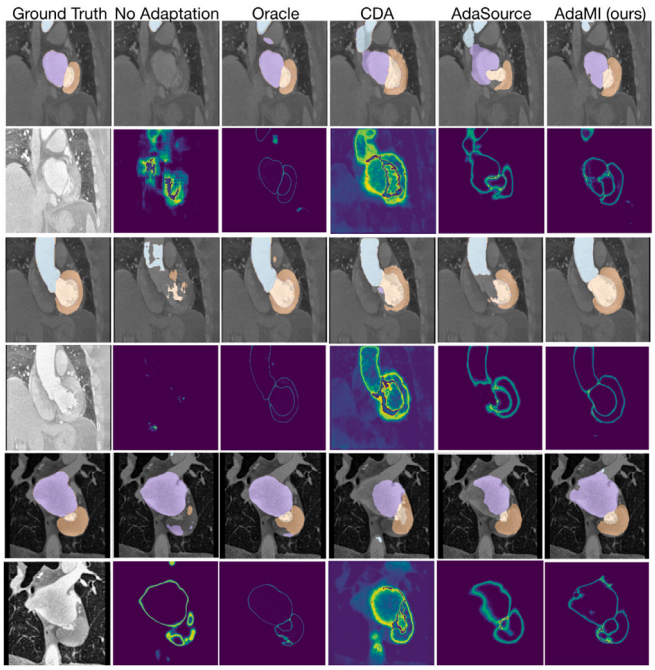

Fig. 9. Qualitative performance on cardiac CT images: examples of the segmentations achieved by our formulation (*AdaMI*), benchmark models in Bateson et al. (2021), Zhang et al. (2019) and lower (*NoAdap*) and upper baselines (*Oracle*). The cardiac structures of MYO, LA, LV and AA are depicted in brown, purple, yellow and blue, respectively.

yet effective estimation of target class-ratios. This finding has great potential value in the medical domain, as prior anatomical knowledge is commonly available, due to conventions in patient position and anatomical similarity (Jurdi et al., 2020). We have, therefore, proposed an effective method to integrate such domain-invariant knowledge, with straightforward extensions in many medical applications. Moreover, the method seems robust enough to adapt to cohorts with possible anatomical variability, i.e. a large shift of class-ratio distributions compared to anatomical reference values (e.g. population-wise differences). Indeed, we show in Table that our model large ambiguities ($\pm 60\%$) on these class-ratios distributions only degrade the performance by up to 15%.

We have also shown that, in the ideal setting when a very precise prior is known, the performance of *AdaMI* is close to full supervision. This suggests that *AdaMI* is able to approach the *"optimal" segmentation given the amount of prior imprecision*. This finding is in line with the recent work of Kervadec et al. (2021), which shows that using a few global shape descriptors as supervision enables performances close to a full pixel-wise supervision. In fact, the class-ratio used in *AdaMI* is based on zero-order shape moments.

We have also demonstrated the superiority of *AdaMI* when compared to our previous *AdaEnt*, which regularizes the class-ratio priors with a steeper loss (Bateson et al., 2020). Indeed, *AdaMI* is able to prevent the under-segmentation regime observed without adaptation, while avoiding the fast convergence to local minima observed with *AdaEnt*. Although convergence and stability are well-known challenges for unsupervised and weakly supervised deep domain adaptation methods, *AdaMI* shows remarkable training stability. On cross-modality spine MRI and cross-site prostate MRI, our method has shown performances on par with several adaptation models that necessitate both the source and target data, such as (Bateson et al., 2021; Zhang et al., 2019). Surprisingly, for the adaptation of MRI-to-CT cardiac images, our model outperforms several recent state-of-the-art adaptation models, such as (Bateson et al., 2021; Zhang et al., 2019; Tsai et al., 2018; Chen et al., 2020; Hoffman et al., 2018; Zhu et al., 2017; Tzeng

et al., 2017; Ganin et al., 2016). This is confirmed qualitatively by our experiments, where the structures of interests are well predicted in all the three applications. In some cases, the segmentation masks are even improved when compared to benchmark adaptation models, despite the lack of source data. These results, therefore, suggest that having access to the source data may not be necessary for domain adaptation.

Extension to 3D: In our experiments on all three applications, the images are 3D volumes. As we have used a standard 2D segmentation network (2D-UNet Ronneberger et al., 2015), we input slices from these 3D volumes for training and inference. However, our method can be extended to be fully-3D; to this end, 3D class-ratio priors should be obtained, to adapt a 3D segmentation network (such as 3D-UNet Çiçek et al., 2016).

Limitations: A limitation of our work is the need for an image-level annotation, compared to fully unsupervised domain adaptation methods. Such annotation for each slice of each volumetric test image in every new target domain can add substantial annotation cost. However, the majority of unsupervised domain adaptation methods use both the source and target data, and are much more complex. Very recent test-time domain adaptation methods such as (He et al., 2020; Karani et al., 2021) also comply with the source-free domain adaptation scenario, but at the cost of an auxiliary branch or additional training tasks in source training phase. Instead, our method tackles the adaptation problem with no alteration in the source training phase, by optimizing a single network, and uses only the target images in the adaptation phase. Importantly, this drastically reduces the computational burden, while easing optimization difficulty, when compared to state-of-the-art domain adaptation models, notably adversarial methods. Indeed, these methods rely on a two-step training of two networks, a discriminator and a segmenter, and a dependency on data from both the source and target domains.

5. Conclusion

Our proposed Source-Free Domain Adaptation (SFDA) tackles a source-free domain adaptation for semantic segmentation, which removes the need for a concurrent access to the source and target data during adaptation. Our approach substitutes the standard supervised loss in the source domain by a direct minimization of the entropy of predictions in the target domain. To prevent trivial solutions, we regularize the entropy loss with a class-ratio prior, which is derived from approximate anatomical knowledge. Unlike recent domain-adaptation techniques, our method tackles domain adaptation without resorting to source data during the adaptation phase, a setting of great value in practice. Interestingly, our formulation achieves a better performance than several state-of-the-art methods which still need access to both source and target data. Our source-free approach has been validated with cross-modality intervertebral disks segmentation, cross-site prostate segmentation and MRI to CT cardiac substructure segmentation. This shows the effectiveness of our prior-aware entropy minimization and that adaptation might not need access to the source data, even when the domain shift is large, as suggested by our experiment on MR to CT cardiac images. Future work will address the integration of other anatomical priors. Our proposed adaptation framework is straightforward to use, drastically reduces the computational burden of the domain adaptation, the optimization complexity, and can be used with any segmentation network architecture.

Declaration of competing interest

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.

Acknowledgments

This work is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), Discovery Grant program, by the The Fonds de recherche du Québec - Nature et technologies (FRQNT), Canada grant, the Canada Research Chair, Canada on Shape Analysis in Medical Imaging, the ETS Research Chair on Artificial Intelligence in Medical Imaging, and NVIDIA, United States with a donation of a GPU. The authors would like to thank the MICCAI 2018 IVDM3Seg, NCI-ISBI 2013, and MICCAI 2017 MMWHS organizers for providing the data.

Appendix A. Estimation of the class-ratio priors from anatomical knowledge

We detail below the estimation of the class-ratio priors for each application. Note that for a structure k, after obtaining the estimated size in mm^2, the class-ratio (i.e. region proportion) τ_k is calculated as: $\tilde{\tau}_k = \frac{size_k}{R_1 * R_2 * \Omega}$, where R_1 and R_2 are the resolution values in the corresponding plane ($R_1 = R_2$ when isotropic) and Ω is the cardinal size of the image. Table A.4 summarizes the estimations obtained for each structure.

IVDM3Seg. Monitoring Lumbar intervertebral disk dimensions is useful treat lumbar spine diseases and for surgical reconstructions. Reference average Lumbar disk height H was available in Bach et al. (2019), and the antero-posterior Lumbar disk diameter D was available in Mirab et al. (2018). The area for one IVD is obtained as $size_{IVD} = \frac{\pi \times H \times D}{4}$, and the final area is $size_{IVDS} = 7 \times size_{IVD}$.

NCI-ISBI13. Prostate volume and dimensions are widely monitored. Reference volume V and height H were taken from (Eri et al., 2002), which measured them through planimetry. We calculated the transverse surface dimension as: $size_{Prostate} = \frac{3V}{2H}$.

MMWHS.[4] **LA** Reference LA area dimensions are readily available as LA area is a useful biomarker in clinical assessment of heart diseases. We used the measurement in Anderson et al. (2005) Table 1, taken at maximum volume (end-systole) in the 4-chamber view[5]; **LV** In O'Dell (2019) Table 3, an estimation of LV area is computed by averaging measurements across 12 long-axis angles.[4] We took the measurement at maximum volume, i.e. end-diastole, to estimate $size_{LV}$; **AA** We used aortic diameters at proximal (p) and distal (d) levels as given in Aronberg et al. (1984), as well as the average AA length (l) provided by the MMWHS organisers[6] to calculate an estimation of the average AA area in a coronal slice as: $size_{AA} = \frac{p+d}{2} * l + \pi * (p/4)^2$. **Myo** The Myo is the structure with the most complicated geometry, thus obtaining an accurate estimation of $size_{Myo}$ is difficult. However, in Støylen et al. (2020), left ventricular myocardial and cavity volumes are available at end-diastole (LVEDV and MVd respectively) and end-systole (LVESV and MVs). We calculate these two ratios: $r_{diastole} = \frac{LVEDV}{MVd}$; $r_{systole} = \frac{LVESV}{MVs}$ and estimate the average Myo area in a coronal slice as: $size_{AA} = \frac{r_{diastole} + r_{systole}}{2} * size_{LV}$.

Appendix B. Link between the loss in *AdaMI* and mutual information maximization

Given the following expression of the mutual information between two random variables X and Y:

$$I(X;Y) = \mathbb{E}_Y \left[\log \mathbb{E}_X[p(Y \mid X)]\right] - \mathbb{E}_{X,Y}[\log p(Y \mid X)]$$

[4] As we used the preprocessed data from Dou et al. (2019), which had performed cropping, zooming and resampling of the slices, we estimated the resolution of these preprocessed slices in the coronal plane as 0.45×0.93 mm/px

[5] Note that these planes are slightly different from the coronal imaging plane of the cardiac slices used in our framework, leading to imprecisions in our estimations.

[6] http://www.sdspeople.fudan.edu.cn/zhuangxiahai/0/mmwhs/data.html

Table A.4
Estimated sizes and class-ratios of structures in the target datasets.

	IVDM3Seg	NCI-ISBI13	MMWHS			
	IVD	Prostate	Myo	LA	LV	AA
$size_k$ (mm2)	2784	2485	1871	2110	1895	1565
$size_k$ (pix)	1782	6095	4428	4996	4487	3706
$\bar{\tau}_k$ (%)	2.72	4.68	6.76	7.62	6.85	5.65

The mutual information between an input image I_t and its softmax predictions P_t can be written as:

$$\mathcal{I}(I_t; P_t) = \mathbb{E}_{P_t}\left[\log \mathbb{E}_{I_t}[p(P_t \mid I_t)]\right] - \mathbb{E}_{I_t,P_t}[\log p(P_t \mid I_t)]$$

And recall that $\mathbb{E}_{I_t}[p(P_t \mid I_t)] = \frac{1}{|\Omega_t|}\sum_{i\in\Omega_t}\mathbf{p}_t(i,\theta) = \hat{\tau}(t,.,\theta)$. Decomposing for each term, and assuming pixel-wise independence of P_t, we obtain:

$$\begin{aligned}\mathbb{E}_{P_t}\left[\log \mathbb{E}_{I_t}[p(P_t \mid I_t)]\right] &= \mathbb{E}_{P_t}\left[\log \hat{\tau}(t,.,\theta)\right] \\ &= \frac{1}{|\Omega_t|}\sum_{i\in\Omega_t}\sum_{k=1}^{K} p_t^k(i,\theta)\log\hat{\tau}(t,k,\theta) \\ &= \sum_{k=1}^{K}\hat{\tau}(t,k,\theta)\log\hat{\tau}(t,k,\theta) = -H\{\hat{\tau}(t,.,\theta)\}\end{aligned}$$

and:

$$\begin{aligned}-\mathbb{E}_{I_t,P_t}[\log p(P_t \mid I_t)] &= -\frac{1}{|\Omega_t|}\sum_{i\in\Omega_t}\sum_{k=1}^{K} p_t^k(i,\theta)\log p_t^k(i,\theta) \\ &= \frac{1}{|\Omega_t|}\sum_{i\in\Omega_t}\ell_{ent}(\mathbf{p}_t(i,\theta))\}\end{aligned}$$

The following identity follows:

$$\mathcal{I}(I_t; P_t) = \underbrace{-H\{\hat{\tau}(t,.,\theta)\}}_{\mathbb{E}_{P_t}\left[\log \mathbb{E}_{I_t}[p(P_t|I_t)]\right]} + \underbrace{\frac{1}{|\Omega_t|}\sum_{i\in\Omega_t}\ell_{ent}(\mathbf{p}_t(i,\theta))}_{-\mathbb{E}_{I_t,P_t}[\log p(P_t|I_t)]}$$

Finally the empirical estimation of the mutual information between a set of input images I_t and their latent label predictions P_t, $t = 1 \dots T$ is given by:

$$\mathcal{I}_\theta = \frac{1}{T}\sum_{t=1}^{T}\mathcal{I}(I_t; P_t) = \frac{1}{T}\sum_{t=1}^{T} -H\{\hat{\tau}(t,.,\theta)\} + \frac{1}{|\Omega_t|}\sum_{i\in\Omega_t}\ell_{ent}(\mathbf{p}_t(i,\theta))$$

References

Anderson, J., Horne, B., Pennell, D., 2005. Atrial dimensions in health and left ventricular disease using cardiovascular magnetic resonance. J. Soc. Cardiovasc. Magn. Reson. 7, 671–675.

Aronberg, D., Glazer, H., Madsen, K., Sagel, S., 1984. Normal thoracic aortic diameters by computed tomography. Comput. Assist. Tomography 8 (2), 247–250, PMID: 6707274.

Bach, K., Ford, J., Foley, R., Januszewski, J., Murtagh, R., Decker, S., Uribe, J.S., 2019. Morphometric analysis of lumbar intervertebral disc height: An imaging study. World Neurosurg. 124, e106–e118.

Bateson, M., Dolz, J., Kervadec, H., Lombaert, H., Ben Ayed, I., 2021. Constrained domain adaptation for image segmentation. IEEE Trans. Med. Imaging 40 (7), 1875–1887.

Bateson, M., Kervadec, H., Dolz, J., Lombaert, H., Ben Ayed, I., 2020. Source-relaxed domain adaptation for image segmentation. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, pp. 490–499.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., Vaughan, J.W., 2010. A theory of learning from different domains. Mach. Learn. 79 (1), 151–175.

Benaim, S., Wolf, L., 2018. One-shot unsupervised cross domain translation. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. NIPS '18, pp. 2108–2118.

Bian, C., Yuan, C., Wang, J., Li, M., Yang, X., Yu, S., Ma, K., Yuan, J., Zheng, Y., 2020. Uncertainty-aware domain alignment for anatomical structure segmentation. Med. Image Anal. 64, 101732.

Billot, B., Greve, D.N., Puonti, O., Thielscher, A., Van Leemput, K., Fischl, B., Dalca, A.V., Iglesias, J.E., 2021. Synthseg: Domain randomisation for segmentation of brain MRI scans of any contrast and resolution. arXiv:2107.09559 [Cs].

Billot, B., Greve, D.N., Van Leemput, K., Fischl, B., Iglesias, J.E., Dalca, A., 2020. A learning strategy for contrast-agnostic MRI segmentation. In: Medical Imaging with Deep Learning. pp. 75–93.

Boudiaf, M., Kervadec, H., Masud, Z.I., Piantanida, P., Ben Ayed, I., Dolz, J., 2021. Few-shot segmentation without meta-learning: A good transductive inference is all you need? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13979–13988.

Chang, W.-L., Wang, H.-P., Peng, W.-H., Chiu, W.-C., 2019. All about structure: Adapting structural information across domains for boosting semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1900–1909.

Chen, C., Dou, Q., Chen, H., Qin, J., Heng, P.A., 2020. Unsupervised bidirectional cross-modality adaptation via deeply synergistic image and feature alignment for medical image segmentation. IEEE Trans. Med. Imaging 39 (7), 2494–2505.

Cheplygina, V., de Bruijne, M., Pluim, J.P.W., 2019. Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. Med. Image Anal. 54, 280–296.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O., 2016. 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In: Proceedings of the International Con- Ference on Medical Image Computing and Computer-Assisted Intervention. MICCAI 2016, Springer International Publishing, Cham, pp. 424–432.

Crammer, K., Kearns, M., Wortman, J., 2007. Learning from multiple sources. In: Advances in Neural Information Processing Systems. pp. 321–328.

Dorent, R., Joutard, S., Shapey, J., Bisdas, S., Kitchen, N., Bradford, R., Saeed, S., Modat, M., Ourselin, S., Vercauteren, T., 2020. Scribble-based domain adaptation via co-segmentation. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 479–489.

Dorent, R., Joutard, S., Shapey, J., Kujawa, A., Modat, M., Ourselin, S., Vercauteren, T., 2021. Inter extreme points geodesics for end-to-end weakly supervised image segmentation. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 615–624.

Dou, Q., Ouyang, C., Chen, C., Chen, H., Glocker, B., Zhuang, X., Heng, P., 2019. PnP-AdaNet: Plug-and-play adversarial domain adaptation network at unpaired cross-modality cardiac segmentation. IEEE Access 7, 99065–99076.

Eri, L.M., Thomassen, H., Brennhovd, B., Håheim, L.L., 2002. Accuracy and repeatability of prostate volume measurements by transrectal ultrasound. Prostate Cancer Prostatic Diseases 5 (4), 273–278.

Ganin, Y., Lempitsky, V., 2015. Unsupervised domain adaptation by backpropagation. In: Bach, F., Blei, D. (Eds.), Proceedings of the 32nd International Conference on Machine Learning. In: Proceedings of Machine Learning Research, vol. 37, PMLR, pp. 1180–1189.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V., 2016. Domain-adversarial training of neural networks. J. Mach. Learn. Res. 17 (1), 2030–2096.

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets. In: Proceedings of Advances in Neural Information Processing Systems. MIT Press, pp. 2672–2680.

Grandvalet, Y., Bengio, Y., 2004. Semi-supervised learning by entropy minimization. In: Advances in Neural Information Processing Systems, Vol. 17.

He, Y., Carass, A., Zuo, L., Dewey, B.E., Prince, J.L., 2020. Self domain adapted network. In: Martel, A.L., Abolmaesumi, P., Stoyanov, D., Mateus, D., Zuluaga, M.A., Zhou, S.K., Racoceanu, D., Joskowicz, L. (Eds.), Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer International Publishing, Cham, pp. 437–446.

He, Y., Carass, A., Zuo, L., Dewey, B.E., Prince, J.L., 2021. Autoencoder based self-supervised test-time adaptation for medical image analysis. Med. Image Anal. 72, 102136.

Hoffman, J., Tzeng, E., Park, T., Zhu, J.-Y., Isola, P., Saenko, K., Efros, A., Darrell, T., 2018. Cycada: Cycle-consistent adversarial domain adaptation. In: Int. Conf. Machine Learning. pp. 1989–1998.

Huo, Y., Xu, Z., Bao, S., Assad, A., Abramson, R.G., Landman, B.A., 2018. Adversarial synthesis learning enables segmentation without target modality ground truth. In: IEEE Int. Symp. on Biomedical Imaging. ISBI, pp. 1217–1220.

Huo, Y., Xu, Z., Moon, H., Bao, S., Assad, A., Moyo, T.K., Savona, M.R., Abramson, R.G., Landman, B.A., 2019. SynSeg-Net: Synthetic segmentation without target modality ground truth. IEEE Trans. Med. Imaging 38 (4), 1016–1025.

Jabi, M., Pedersoli, M., Mitiche, A., Ayed, I.B., 2021. Deep clustering: On the link between discriminative models and K-means. IEEE Trans. Pattern Anal. Mach. Intell. 43 (6), 1887–1896.

Javanmardi, M., Tasdizen, T., 2018. Domain adaptation for biomedical image segmentation using adversarial training. In: IEEE Int. Symp. on Biomedical Imaging. ISBI, pp. 554–558.

Jia, Z., Huang, X., Eric, I., Chang, C., Xu, Y., 2017. Constrained deep weak supervision for histopathology image segmentation. IEEE Trans. Med. Imaging 36 (11), 2376–2388.

Jurdi, R.E., Petitjean, C., Honeine, P., Cheplygina, V., Abdallah, F., 2020. High-level prior-based loss functions for medical image segmentation: A survey. arXiv:2011.08018.

Kamnitsas, K., Baumgartner, C.F., Ledig, C., Newcombe, V.F.J., Simpson, J.P., Kane, A.D., Menon, D.K., Nori, A.V., Criminisi, A., Rueckert, D., Glocker, B., 2017. Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. In: Information Processing in Medical Imaging. IPMI, pp. 597–609.

Karani, N., Erdil, E., Chaitanya, K., Konukoglu, E., 2021. Test-time adaptable neural networks for robust medical image segmentation. Med. Image Anal. 68, 101907.

Kervadec, H., Bahig, H., Létourneau-Guillon, L., Dolz, J., Ayed, I.B., 2021. Beyond pixel-wise supervision for segmentation: A few global shape descriptors might be surprisingly good!. In: Proc. Conf. Medical Imaging with Deep Learning. MIDL, pp. 354–368.

Kervadec, H., Dolz, J., Granger, É., Ben Ayed, I., 2019a. Curriculum semi-supervised segmentation. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Springer, pp. 568–576.

Kervadec, H., Dolz, J., Tang, M., Granger, E., Boykov, Y., Ayed, I.B., 2019b. Constrained-CNN losses for weakly supervised segmentation. Med. Image Anal. 54, 88–99.

Kervadec, H., Dolz, J., Wang, S., Granger, E., ben Ayed, I., 2020. Bounding boxes for weakly supervised segmentation: Global constraints get close to full supervision. In: Proc. Conf. Medical Imaging with Deep Learning. MIDL, pp. 365–381.

Khan, S., Shahin, A.H., Villafruela, J., Shen, J., Shao, L., 2019. Extreme points derived confidence map as a cue for class-agnostic interactive segmentation using deep neural network. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 66–73.

Kingma, D., Ba, J., 2014. Adam: A method for stochastic optimization. In: Int. Conf. on Learning Representations. ICLR.

Krause, A., Perona, P., Gomes, R., 2010. Discriminative clustering by regularized information maximization. In: Advances in Neural Information Processing Systems, Vol. 23. Curran Associates, Inc..

Li, Y., Yuan, L., Vasconcelos, N., 2019. Bidirectional learning for domain adaptation of semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6929–6938.

Liang, J., Hu, D., Feng, J., 2020. Do we really need to access the source data? Source hypothesis transfer for unsupervised domain adaptation. In: Proc. 37th Int. Conf. on Machine Learning. In: Proc. of Machine Learning Research, vol. 119, PMLR, pp. 6028–6039.

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A., van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. Med. Image Anal. 42, 60–88.

Liu, Q., Dou, Q., Heng, P.-A., 2020. Shape-aware meta-learning for generalizing prostate MRI segmentation to unseen domains. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, pp. 475–485.

Long, M., Cao, Y., Wang, J., Jordan, M.I., 2015. Learning transferable features with deep adaptation networks. In: Proceedings of the 32nd International Conference on International Conference on Machine Learning - Vol. 37. ICML '15, pp. 97–105.

Luo, Y., Zheng, L., Guan, T., Yu, J., 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. pp. 2502–2511.

Mirab, S.M.H., Barbarestani, M., Tabatabaei, S.M., Shahsavari, S., Minaeii Zangi, M.B.a., 2018. Measuring dimensions of lumbar intervertebral discs in normal subjects. Anatomical Sci. J. 15 (1).

Morerio, P., Cavazza, J., Murino, V., 2018. Minimal-entropy correlation alignment for unsupervised deep domain adaptation. In: Int. Conf. on Learning Representations. ICLR.

Nath Kundu, J., Venkat, N., Rahul, M.V., Venkatesh Babu, R., 2020. Universal source-free domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4543–4552.

O'Dell, W.G., 2019. Accuracy of left ventricular cavity volume and ejection fraction for conventional estimation methods and 3D surface fitting. J. Am. Heart Assoc. 8 (6).

Ouyang, X., Xue, Z., Zhan, Y., Zhou, X.S., Wang, Q., Zhou, Y., Wang, Q., Cheng, J.-Z., 2019. Weakly supervised segmentation framework with uncertainty: A study on pneumothorax segmentation in chest X-ray. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 613–621.

Pan, S.J., Yang, Q., 2010. A survey on transfer learning. IEEE Trans. Knowl. Data Eng. 22 (10), 1345–1359.

Patel, G., Dolz, J., 2022. Weakly supervised segmentation with cross-modality equivariant constraints. Med. Image Anal. 77, 102374.

Paul, S., Tsai, Y.-H., Schulter, S., Roy-Chowdhury, A.K., Chandraker, M., 2020. Domain adaptive semantic segmentation using weak labels. In: European Conference on Computer Vision. ECCV, Springer, pp. 571–587.

Rajchl, M., Lee, M.C., Oktay, O., Kamnitsas, K., Passerat-Palmbach, J., Bai, W., Damodaram, M., Rutherford, M.A., Hajnal, J.V., Kainz, B., et al., 2016. Deepcut: Object segmentation from bounding box annotations using convolutional neural networks. IEEE Trans. Med. Imaging 36 (2), 674–683.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional networks for biomedical image segmentation. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, pp. 234–241.

Sankaranarayanan, S., Balaji, Y., Castillo, C., Chellappa, R., 2018. Generate to adapt: Aligning domains using generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8503–8512.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 618–626.

Støylen, A., Dalen, H., Molmen, H.E., 2020. Left ventricular longitudinal shortening: Relation to stroke volume and ejection fraction in ageing, blood pressure, body size and gender in the HUNT3 study. Open Heart 7 (2), e001243.

Sun, Y., Wang, X., Liu, Z., Miller, J., Efros, A., Hardt, M., 2020. Test-time training with self-supervision for generalization under distribution shifts. In: III, H.D., Singh, A. (Eds.), Proceedings of the 37th International Conference on Machine Learning. In: Proceedings of Machine Learning Research, vol. 119, PMLR, pp. 9229–9248.

Tang, M., Perazzi, F., Djelouah, A., Ben Ayed, I., Schroers, C., Boykov, Y., 2018. On regularized losses for weakly-supervised CNN segmentation. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 507–522.

Tsai, Y.-H., Hung, W.-C., Schulter, S., Sohn, K., Yang, M.-H., Chandraker, M., 2018. Learning to adapt structured output space for semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7472–7481.

Tulder, G.v., Bruijne, M.d., 2016. Representation learning for cross-modality classification. In: Medical Computer Vision and Bayesian and Graphical Models for Biomedical Imaging. Springer, pp. 126–136.

Tzeng, E., Hoffman, J., Darrell, T., Saenko, K., 2015. Simultaneous deep transfer across domains and tasks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4068–4076.

Tzeng, E., Hoffman, J., Saenko, K., Darrell, T., 2017. Adversarial discriminative domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2962–2971.

Varsavsky, T., Orbes-Arteaga, M., Sudre, C., Graham, M., Nachev, P., Cardoso, M.J., 2020. Test-time unsupervised domain adaptation. In: Martel, A., Abolmaesumi, P., Stoyanov, D., Mateus, D., Zuluaga, M., Zhou, S.K., Racoceanu, D., Joskowicz, L. (Eds.), Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. In: Lecture Notes in Computer Science, Springer Science and Business Media Deutschland GmbH, Germany, pp. 428–436.

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

Wachinger, C., Reuter, M., 2016. Domain adaptation for Alzheimer's disease diagnostics. NeuroImage 139, 470–479.

Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T., 2021. Tent: Fully test-time adaptation by entropy minimization. In: International Conference on Learning Representations.

Wu, K., Du, B., Luo, M., Wen, H., Shen, Y., Feng, J., 2019. Weakly supervised brain lesion segmentation via attentional representation learning. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 211–219.

Wu, X., Zhang, S., Zhou, Q., Yang, Z., Zhao, C., Latecki, L.J., 2020. Entropy minimization vs. Diversity maximization for domain adaptation. arXiv:2002.01690.

Zhang, Y., David, P., Foroosh, H., Gong, B., 2019. A curriculum domain adaptation approach to the semantic segmentation of urban scenes. IEEE Trans. Pattern Anal. Mach. Intell. 42 (8), 1823–1841.

Zhang, Y., Miao, S., Mansi, T., Liao, R., 2018a. Task driven generative modeling for unsupervised domain adaptation: Application to X-ray image segmentation. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 599–607.

Zhang, Y., Qiu, Z., Yao, T., Liu, D., Mei, T., 2018b. Fully convolutional adaptation networks for semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6810–6818.

Zhao, H., Li, H., Maurer-Stroh, S., Guo, Y., Deng, Q., Cheng, L., 2018. Supervised segmentation of un-annotated retinal fundus images by synthesis. IEEE Trans. Med. Imaging 38 (1), 46–56.

Zhou, Y., Li, Z., Bai, S., Chen, X., Han, M., Wang, C., Fishman, E., Yuille, A., 2019. Prior-aware neural network for partially-supervised multi-organ segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10672–10681.

Zhu, J., Park, T., Isola, P., Efros, A.A., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision. ICCV, pp. 2242–2251.

Zhuang, X., Li, L., Payer, C., Štern, D., Urschler, M., Heinrich, M.P., Oster, J., Wang, C., Smedby, Ö., Bian, C., Yang, X., Heng, P.-A., Mortazi, A., Bagci, U., Yang, G., Sun, C., Galisot, G., Ramel, J.-Y., Brouard, T., Tong, Q., Si, W., Liao, X., Zeng, G., Shi, Z., Zheng, G., Wang, C., MacGillivray, T., Newby, D., Rhode, K., Ourselin, S., Mohiaddin, R., Keegan, J., Firmin, D., Yang, G., 2019. Evaluation of algorithms for multi-modality whole heart segmentation: An open-access grand challenge. Med. Image Anal. 58, 101537.

Zou, Y., Yu, Z., Kumar, B.V.K.V., Wang, J., 2018. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 289–305.