On Direct Distribution Matching for Adapting Segmentation Networks

Georg Pichler[1], Jose Dolz[2], Ismail Ben Ayed[2], and Pablo Piantanida[1,3]

[1]CentraleSupélec-CNRS-Université Paris Sud, 3 rue Joliot-Curie, Gif-sur-Yvette, France
[2]ÉTS, 1100 Notre Dame St W, Montreal, Canada
[3]Montreal Institute for Learning Algorithms (Mila), Université de Montréal, QC, Canada
{georg.pichler, pablo.piantanida}@l2s.centralesupelec.fr,
{Jose.Dolz,Ismail.BenAyed}@etsmtl.ca

April 5, 2019

Minimization of distribution matching losses is a principled approach to domain adaptation in the context of image classification. However, it is largely overlooked in adapting segmentation networks, which is currently dominated by adversarial models. We propose a class of loss functions, which encourage direct kernel density matching in the network-output space, up to some geometric transformations computed from unlabeled inputs. Rather than using an intermediate domain discriminator, our direct approach unifies distribution matching and segmentation in a single loss. Therefore, it simplifies segmentation adaptation by avoiding extra adversarial steps, while improving both the quality, stability and efficiency of training. We juxtapose our approach to state-of-the-art segmentation adaptation via adversarial training in the network-output space. In the challenging task of adapting brain segmentation across different magnetic resonance images (MRI) modalities, our approach achieves significantly better results both in terms of accuracy and stability.

1. Introduction

Semantic segmentation is of pivotal importance towards high-level understanding of image content, which is useful in a breadth of application areas, from autonomous driving to health care, for instance. Particularly, in medical imaging, segmentation facilitates clinical tasks, including

1

disease diagnosis, treatment and follow-up, among others. Modern medical segmentation approaches rely on deep learning techniques, which have demonstrated outstanding performances in a breadth of applications [7, 8, 24]. Despite their success, generalization of trained models to new scenarios may be hampered if the gap between data distributions across domains is large. A trivial solution to address this issue would be to re-annotate images from different domains and re-train or fine-tune the deep models. Nevertheless, obtaining such massive amounts of labeled data is a cumbersome process which, for some applications, may require user expertise, resulting in a prohibitive and unrealistic solution.

To tackle this problem, unsupervised domain adaptation (UDA) techniques have been widely investigated. These methods aim at learning robust classifiers, or other predictors, in the presence of a *shift* between source and target distributions when the target data is unlabeled. An important body of recent literature formulates UDA as a domain divergence minimization. In this scenario, the goal is typically to minimize the discrepancy between distributions across domains at the inputs [1, 3, 15, 32, 34, 42] or at intermediate-feature levels [11, 12, 21, 25, 26, 27, 39], while leveraging labeled source examples to retain discriminative power on the feature space. Generative pixel-level techniques align the image appearance between domains, so that the target data 'style' is transferred to source data, or vice-versa [1, 3, 32, 35, 45]. Then, supervised learning with the newly generated synthetic data is performed to train a segmentation network on transformed data. A downside of these approaches is that they perform satisfactorily only for small images and narrow domain shifts, which limits their applicability. Feature-level UDA models follow a similar distribution alignment but in the feature space instead [10, 11, 21, 26, 39]. Within the current paradigm of learning domain-invariant representations, domain adversarial training [11, 39] and maximum mean discrepancies (MMD) [26, 37, 43] have become very popular choices.

For semantic segmentation problems, adversarial training models [14] are currently dominating the recent literature [6, 5, 9, 21, 16, 17, 33, 38, 40]. Such models alternate the training of two networks: a discriminator that learns a decision boundary between source and target features and a segmentation network that uses the learned decision boundary, thereby matching feature distribution across domains. Some other approaches rely on generative networks, which yield target images conditioned on the source, or vice-versa, aligning both domains at the pixel level [2, 18, 29, 34, 46, 47].

While adversarial training achieved outstanding performances in image classification, our numerical evidence and rational intuition suggest that it may not be generally suitable for segmentation tasks. First, in segmentation, learning a discriminator boundary is much more complex than classification as it solves for predictions in an exponentially large label space. Intuitively, a large label space implies large spaces of possible solutions for discriminator boundaries and target predictions, both of which are latent. Therefore, as we will see later in our experiments, alternating both adversarial and prediction tasks in segmentation might cause much more significant training instabilities than in image classification tasks. Furthermore, most of adversarial methods perform feature adaptation at many levels of abstraction. However, a large label space makes the hypothesis that source and target domains share the same multi-level feature representations less likely to be valid. While the inputs can be significantly different from one domain to another, the output (label) space in semantic segmentation convey very rich information related

to the spatial layout and local context, which is shared across domains (See Fig. 1). Inspired by this observation, Tsai *et at.*[38] proposed adversarial training in the output (softmax segmentation) space, achieving better performance than features-matching approaches on the Cityscapes dataset. Leveraging this information is even more meaningful in medical images, where label (output) statistics remain domain-independent, up to geometric transformations, despite significant differences in image inputs across domains. Nevertheless, following the trend in UDA approches for natural image segmentation, adversarial learning has become the *de facto* choice in medical image segmentation [4, 9, 13, 19, 21, 45, 47].

It is worth mentioning that some recent natural image segmentation works [44, 48] pointed out that adversarial models for classification do not translate well to segmentation. These studies showed that similar or better performances can be achieved by other alternatives. The authors of [48] tackled the problem with self-training, which generates masks of unlabeled target images via the network's own predictions. In a different approach, Zhang *et at.*[44] adopted a curriculum learning approach, enforcing consistency of global label distributions such as region proportions. Finally, in the context of classification, the authors of [36] pointed out that UDA based on adversarial training may not be sufficient for models with high capacity, which is the case in segmentation. They empirically proved that, for sufficiently deep architectures, jointly achieving small source generalization error and feature divergence does not imply high accuracy on the target task.

Minimization of distribution matching losses is a principled approach to domain adaptation in the context of image classification, e.g., MDD [26, 37, 43]. However, it is largely overlooked in adapting segmentation networks, which is currently dominated by adversarial models. We propose a class of loss functions, which encourage direct kernel density matching in the network-output space, up to some geometric transformations computed from unlabeled inputs. Rather than using an intermediate domain discriminator, our direct approach unifies distribution matching and segmentation in a single loss. Therefore, it simplifies segmentation adaptation by avoiding extra adversarial steps, while improving quality, stability and efficiency of training. We juxtapose our approach to the state-of-art segmentation method in [38], which performs distribution matching with a two-step adversarial learning in the network-output space. In the challenging task of adapting brain segmentation across different magnetic resonance images (MRI) modalities, our approach achieves significantly better performance than adversarial output adaption, both in terms of accuracy and stability.

2. Formulation

Consider an unsupervised domain-adaptation setting with two distinct subsets: $\mathcal{L} = \{(X_i, Y_i)\}_{i=1,\ldots,n}$ contains labeled source-domain images X_i and the corresponding ground-truth segmentations Y_i, and $\mathcal{U} = \{(X_i, X_i')\}_{i=n+1,\ldots,n+m}$ contains *unlabeled* image pairs, each involving a source image X_i and a target image X_i'. For each labeled source image $X_i : \Omega \subset \mathbb{R}^{2,3} \to \mathbb{R}$, $i = 1, \ldots, n$, the ground-truth labeling $Y_i \in \{0,1\}^{L \times |\Omega|}$ is a matrix whose columns are binary vectors, encoding the assignment of pixel $p \in \Omega$ to one of L classes (segmentation regions): $\mathbf{y}_i(p) = (y_i(1,p), \ldots, y_i(L,p)) \in \{0,1\}^L$. For any image X, let $\mathbf{s}_\theta(p, X) =$
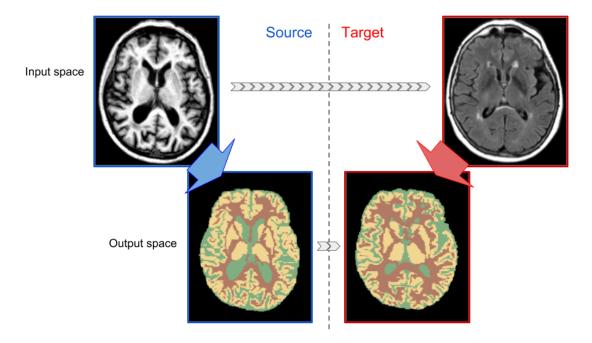
Figure 1: The stat-of-the-art domain adaptation method in [38] proposed adversarial training in the output (softmax segmentation) space, achieving better performance than features-matching approaches. This is motivated by the fact that input images can be significantly different from one domain to another, whereas the output (label) space in semantic segmentation is shared across domains and conveys very rich information related to the spatial layout and local context. Leveraging this information is even more meaningful in medical images, where label (output) statistics remain domain-independent, up to geometric transformations, despite significant differences in image inputs across domains.

(a) Adverserial (Discriminator)   (b) Adverserial (Segmenter)   (c) Direct Distribution Matching
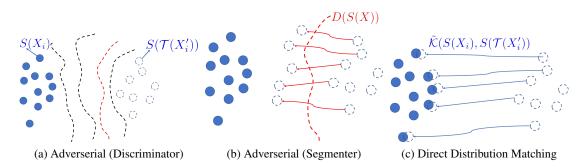
Figure 2: A conceptual juxtaposition of adversarial training in the network-output space [38] (2a and 2b) and our direct kernel density matching (2c).

$\big(s_\theta(1, p, X), \ldots, s_\theta(L, p, X)\big) \in [0, 1]^L$ denote the probability vector of softmax outputs for

Function	$-\mathcal{D}(\mathbf{s}, \mathbf{s}')$
RBF kernel	$-\exp(-\|\mathbf{s} - \mathbf{s}'\|^2/\sigma)$
Bhattacharyya kernel	$-\sqrt{\mathbf{s}^t \mathbf{s}'}$
KL divergence	$\mathbf{s}^t \ln \frac{\mathbf{s}}{\mathbf{s}'}$
Squared Euclidean distance	$\|\mathbf{s} - \mathbf{s}'\|^2$

Table 1: Examples of choices for pairwise discrepancy functions $\mathcal{D}(\mathbf{s}, \mathbf{s}')$ for probability simplex vectors $\mathbf{s}$ and $\mathbf{s}'$, including some kernel-based functions $\mathcal{D} = -\mathcal{K}$. For RBF (Gaussian) kernel, σ is the the kernel width. Superscript t denotes transposition.

pixel p, with θ the trainable parameters of the network. For the sake of simplicity, we will omit the subscript θ in the following.

We propose to minimize the following loss function:

$$\mathcal{F}(\theta) = \sum_{i=1}^{n} \sum_{p \in \Omega} \mathcal{H}\big(\mathbf{y}_i(p), \mathbf{s}(p, X_i)\big)$$
$$+ \lambda \sum_{i=n+1}^{n+m} \sum_{p \in \Omega} \mathcal{D}\big(\mathbf{s}(p, X_i), \mathbf{s}(p, \mathcal{T}(X_i'))\big), \tag{1}$$

where

- $\mathcal{D}(\mathbf{s}, \mathbf{s}')$ evaluates the discrepancy between two probability distributions $\mathbf{s}$ and $\mathbf{s}'$, e.g., the KL divergence. It is possible to choose $\mathcal{D}(\cdot, \cdot) = -\mathcal{K}(\cdot, \cdot)$, with some kernel function $\mathcal{K}$ defined over a pair of probability vectors, e.g., a probability product kernel [20] $\mathcal{K}(\mathbf{s}, \mathbf{s}') = \sum_{l=1}^{L} s(l)^\rho s'(l)^\rho$, with $s(l)$ denoting the lth component of $\mathbf{s}$ and[1] $\rho \in]0, 1]$. Table 1 lists a few choices of $\mathcal{D}$ including some kernel functions.

- $\mathcal{T}$ is a geometric transformation, which aligns pairs of unlabeled images, for instance, using a standard automatic cross-modality registration algorithm [30].

- $\mathcal{H}$ denotes standard cross-entropy loss for labeled source-domain images: $\mathcal{H}(\mathbf{y}_i(p), \mathbf{s}(p, X_i)) = -\mathbf{y}_i^t(p) \ln \mathbf{s}(p, X_i)$, with superscript t denoting transposition.

- λ is a non-negative multiplier.

The second term in our model in (1) encourages the density of network outputs (softmax segmentations) in the target domain to closely match the density of those in the source domain. Our loss in (1) encourages direct kernel density matching in the network-output space. Fig. 2 juxtaposes conceptually our direct matching (Fig. 2c) to the state-of-art method in [38], which pursues a two-step adversarial learning in the network-output space (Figs. 2a and 2b), so as to achieve the same goal as our loss: matching the source and target distributions of label predictions. The data points in the Figure depict networks outputs (softmax segmentations) in $\{0, 1\}^{L \times |\Omega|}$, with the blue points corresponding to the source and dashed points to the target.

The model in [38] alternates the training of two networks: (i) a discriminator (Fig. 2a), which learns a decision boundary $D(S(X))$ that distinguishes between source and target outputs; and

---
[1] Notice that $\rho = \frac{1}{2}$ corresponds to the well-known Bhattacharyya coefficient.

(ii) a segmentation network that uses the learned decision boundary, thereby encouraging the target outputs to be similar to those in the source (Fig. 2b). Rather than using an intermediate domain discriminator, our direct method (Fig. 2c) unifies distribution matching and segmentation in a single loss. Therefore, it simplifies segmentation adaptation by avoiding extra adversarial steps, while improving both the quality, stability and efficiency of training. While adversarial training achieved outstanding performances in image classification, our numerical evidence and intuition suggest that it may not be suitable for segmentation, in which case learning a discriminator boundary is much more complex as it solves for predictions in an exponentially large label space. In fact, intuitively, a large label space implies large spaces of possible solutions for discriminator boundaries and target predictions, both of which are latent; see dashed boundaries and data points in Fig. 2a. Therefore, alternating both adversarial and prediction tasks in segmentation might cause much more significant instabilities than in image classification tasks, as we will see later in our experiments. Another important difference between our approach and adversarial training is that we account for the fact that target and source label predictions are similar up to some geometric transformations. Such a prior information is very common and useful in medical imaging problems but adversarial approaches do not have mechanisms to take advantage of it.

3. Experiments

In this section we describe and present the numerical experiments and results to validate the proposed unsupervised domain adaptation approach. Particularly, we evaluated our approach in the challenging task of brain tissue segmentation on magnetic resonance imaging (MRI). First, we show the performance of our method when it is trained and tested on different image protocols and scans across modalities are perfectly aligned. Then, we evaluate the impact of not having images perfectly aligned, which corresponds to having images from different subjects in the second term of equation (1). In these experiments, we compare the proposed method with AdaptSegNet [38]. And last, we investigated the effect of employing different kernels to evaluate the discrepancy between the source and target probability distributions in (1).

3.1. Experimental details

3.1.1. Dataset

We performed numerical studies in two public segmentation benchmarks: MRBrainS2013 [28] and iSEG2017 [41] challenge. The MRBrainS dataset contains 5 training scans with ground truth and 15 unlabeled scans of adult brains for testing. On the other hand, the iSEG dataset focus on infant brains and it is composed by 10 training and 13 testing scans. We tested our algorithm on the T1 and T2-FLAIR modalities of MRBrainS and T1, T2 of iSEG. In both cases, the segmentation task consists of finding a pixel-wise classification of white matter (WM), gray matter (GM) and cerebrospinal fluid (CSF).

Original T2 images from the iSEG Challenge were resampled into an isotropic $1 \times 1 \times 1\mathrm{mm}^3$ resolution and then aligned onto their corresponding T1 images with a simple affine registration

method. The sequences from the MRBrains Challenge were also aligned by rigid registration, using the Elastix software [23].

3.1.2. Training

We first evaluate our approach when images across domains are perfectly aligned. In this scenario, we assume that the training data $(\mathcal{L}, \mathcal{U})$ consists of two distinct subsets, where $\mathcal{L} = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ contains images X_i from the source domain with their corresponding ground truth y_i, and $\mathcal{U} = \{(X_{n+1}, X'_{n+1}), \dots, (X_{n+m}, X'_{n+m})\}$ contains unlabeled pairs of aligned source and target data, respectively. The calculation of the loss for one batch is detailed in Algorithm 1, where the permutation $\pi = \mathrm{id}$ denotes the identity. In order to test performance when the assumption of alignment between X_i and X'_i is violated, we also perform experiments with shuffled target data, by passing a random permutation π of $\{1, \dots, n\}$ to Algorithm 1. By performing preliminary experiments, we found that the choice of distance functions $\mathcal{D}$ does not significantly alter performance (cf. Section 3.2.4). We performed all experiments with squared Euclidean distance $\mathcal{D}(\mathbf{s}, \mathbf{s}') = \|\mathbf{s} - \mathbf{s}'\|^2$.

Algorithm 1: Computing the loss for one random batch. The function RandomSample($\mathcal{A}, k$) selects k elements from the set $\mathcal{A}$ uniformly, independently at random, without replacement.

Input: batch size t, distance function $\mathcal{D}$, permutation π of $\{1, \dots, m\}$, network $\mathbf{s}(\cdot, \cdot)$, transformation $\mathcal{T}$, Lagrange multiplier λ
Data: $\mathcal{L} = \{(X_1, Y_1), \dots, (X_n, Y_n)\}, \mathcal{U} = \{(X_{n+1}, X'_{n+1}), \dots, (X_{n+m}, X'_{n+m})\}$
Output: loss L for one random batch
begin
$\quad \mathcal{I}_0 \leftarrow$ RandomSample$(\{1, \dots, n\}, t)$
$\quad \mathcal{I}_1 \leftarrow$ RandomSample$(\{1, \dots, m\}, t)$
$\quad L \leftarrow 0$
$\quad$**for** $i \in \mathcal{I}_0$ **do**
$$L \leftarrow L - \sum_{p \in \Omega} \sum_{l=1}^{L} y_i(l, p) \cdot \log s(l, p, X_i)$$
$\quad$**for** $i \in \mathcal{I}_1$ **do**
$$L \leftarrow L + \lambda \sum_{p \in \Omega} \mathcal{D}\big(\mathbf{s}(p, X_{n+i}), \mathbf{s}(p, \mathcal{T}(X'_{n+\pi(i)}))\big)$$

Due to the limited size of the training set, we employed a leave-one-out-cross-validation strategy, where $n-1$ images were used for training, leaving the remaining image for evaluation. Furthermore, the unlabeled testing sequences were employed in the unsupervised term of Eq. (1).

Each experiment was performed three times with different evaluation data and the average as well as the empirical standard deviation over these three runs is reported.

3.1.3. Baselines

Lower and upper bounds. In order to evaluate the impact of the adaptation approaches we trained the segmentation network in a supervised manner on the source and target data, providing a lower and upper bound for the expected unsupervised domain adaptation results. While network trained on source images is referred to as *No adaptation*, the segmentation network directly trained on the target domain is referred to as the *oracle*.

Adversarial state-of-the-art approach. In addition, we compare the proposed approach with the adversarial method proposed in [38]. For a fair comparison, we used the same segmentation network for the proposed and the adversarial approach. Furthermore, we kept the discriminator network architecture, but set the strides of two of the five convolutional layers to one due to a different input size for the discriminator network. The Lagrange multiplier for training the segmentation network was chosen to be $\lambda_{\mathrm{adv}} = 0.1$.

3.1.4. Implementation details

We used a slightly modified UNet [31] for the segmentation task, operating on 2D slices. Particularly, the employed network follows the original implementation [31], but the depth is reduced by one, i.e., max-pool is performed only three times instead of four. The input size is $(120, 120)$, resulting in a lowest resolution of $(11, 11)$ and an output size of $(28, 28)$. We used ReLU activation functions and did not include dropout, to avoid any regularization that does not originate from our proposed domain adaptation strategy. The Lagrange multiplier $\lambda = 0.1$ was equal for all experiments. The input images are sliced along the z-axis to provide the input for the segmentation network.

For the adversarial approach of [38], we used the same segmentation network (U-Net) as described above to facilitate a fair comparison. To keep matters simple, we chose the "single-level" strategy, performing the domain adaptation only on the output layer. We chose the same discriminator model as in [38]. However, to avoid collapse of the discriminator output due to the small input size $(28, 28)$, we changed the stride of the third an fourth 2D convolution layer from two to one.

Implementation was done in TensorFlow, and experiments were run on a server equipped with a NVidia Titan V GPU with 12 GB memory. For all networks, we employed the Adam [22] optimizer with learning rate $\mathrm{lr} = 0.0001$ and a batch size of 32. The code is publicly available at https://github.com/anonymauthor/DDMSegNet

3.1.5. Evaluation

Our evaluation involves a quantitative and a qualitative component. In terms of quantitative evaluation, we resorted to the common Dice coefficient, widely employed in medical image segmentation, to compare the performance of the different methods introduced in Section 3.1.3. The Dice coefficient can be expressed as:

$$\text{DICE}(\hat{Y}, Y) = \frac{2|\hat{Y} \cap Y|}{|\hat{Y}| + |Y|}. \tag{2}$$

Since the images on the datasets are actually volumetric data, the reported dice values are computed over the 3D entire volume.

			Mean Dice			
			Oracle	No adaptation	AdaptSegNet [38]	Proposed
Source	Target		Target$\longrightarrow$Target	Source$\longrightarrow$Target	Source$\longrightarrow$Target	Source$\longrightarrow$Target
MRB (T1)	MRB (T2-FLAIR)	GM	78.67 ± 2.27	49.94 ± 4.57	56.63 ± 3.34	78.25 ± 2.57
		WM	82.44 ± 1.32	21.01 ± 4.58	60.09 ± 3.84	82.74 ± 0.89
		CSF	74.06 ± 1.45	55.89 ± 2.45	61.23 ± 3.84	73.41 ± 1.25
		Mean	78.39 ± 1.51	42.28 ± 0.32	59.32 ± 1.45	78.13 ± 1.36
MRB (T2-FLAIR)	MRB (T1)	GM	85.44 ± 1.53	21.64 ± 5.48	75.35 ± 0.99	84.41 ± 2.10
		WM	89.82 ± 0.71	18.72 ± 9.10	81.73 ± 0.48	88.83 ± 0.79
		CSF	81.14 ± 1.69	47.33 ± 5.27	69.74 ± 1.66	77.78 ± 0.64
		Mean	85.47 ± 0.64	29.23 ± 5.28	75.61 ± 0.26	83.67 ± 1.16
iSEG (T1)	iSEG (T2)	GM	76.78 ± 1.10	58.01 ± 2.09	66.07 ± 1.22	74.12 ± 0.51
		WM	70.93 ± 1.23	40.22 ± 4.46	58.98 ± 4.01	66.49 ± 1.08
		CSF	84.93 ± 0.88	53.03 ± 2.08	77.35 ± 2.98	83.70 ± 1.04
		Mean	77.55 ± 0.87	50.42 ± 2.82	67.47 ± 2.45	74.77 ± 0.61
iSEG (T2)	iSEG (T1)	GM	81.44 ± 0.95	71.59 ± 0.03	72.83 ± 0.31	77.57 ± 0.34
		WM	76.52 ± 2.50	59.68 ± 2.65	65.57 ± 2.18	69.74 ± 1.77
		CSF	89.14 ± 0.13	72.80 ± 1.67	79.42 ± 4.16	87.05 ± 0.09
		Mean	82.36 ± 1.17	68.03 ± 0.49	72.61 ± 1.71	78.12 ± 0.59

Table 2: Domain adaptation results on MRBrains and iSEG dataset, showing the mean Dice coefficient over the three classes (i.e., GM, WM and CSF).

3.2. Results

3.2.1. Precise image alignment.

Table 2 reports the class-specific and mean results when pairs of unlabeled images are perfectly aligned (permutation $\pi = \text{id}$ in Algorithm 1). Looking at the results achieved by the oracle we can observe that without any adaptation strategy the performance dramatically drops, particularly for the WM. The adaptation strategy proposed in [38] is able to infer target domain information during learning and recover segmentation performance. For example, when shifting from T1 to T2 AdaptSegNet improves the mean performance by at least 17%with respect to the *No adaptation* network in both MRBrains and iSEG images. Despite this improvement, there is still a considerable gap compared to the oracle, with a difference of 10.1% and 19.1% for

iSEG and MRBrains, respectively. On the other hand, the increased performance achieved by our method is more pronounced, getting closer to the performance of the oracle. Particularly, in all the four settings, differences with respect to training the network on target images and our method are in the range between $0.3 - 4.2\%$. Furthermore, in most cases, the standard deviation is largely decreased by employing the proposed approach rather than the adversarial method. Another interesting finding when independently analyzing the class-specific results is that the proposed method reliably follows the behaviour of the oracle. For each of the four analyzed settings, the class segmentation rank for both oracle and proposed approach remains the same.

Qualitative results of these models are depicted in Figure 3. Specifically, cross-sectional 2D MRI images of two given patients are shown, for both source and target domains, along with the corresponding ground truth and segmentation masks obtained by the different models. We can observe that if no adaptation method is applied, the model trained on the source domain completely fail to segment the target image. Including an adaptation adversarial module visually improves the segmentation, which aligns with the numerical values reported in Table 2. Having a closer look to the AdaptSegNet segmentation we observe that while the CSF (in brown) seems to correlate with the ground truth, both white and gray matter (in yellow and green, respectively) only capture global information, being imprecise in local details. This can be due to the fact that appearance of this particular structure remains similar across domains, whereas intensity distribution of white and gray matter highly differ between source and target domains. Indeed, this observation also holds for the *No adaptation* setting, where CSF segmentation obtains the best performance for domain adaptation on MRBrains. Contrary, the proposed direct distribution matching method is able to correctly capture differences between images, satisfactorily adapting both domains.
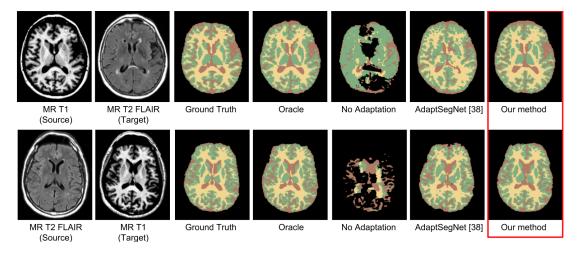

Figure 3: Visual results for two MRBrains subjects achieved by the different models in the case of adapting a MR T1-trained model to MR T2 images (*top*), and a MR T2-trained model to MR T1 images (*bottom*). These images were randomly selected from one of the three runs.

3.2.2. Sensitivity to image disalignment.

In our first experiments we used perfectly aligned images, but different protocols as source and target class. In order to test the performance when the assumption of alignment between X_i and X_i' is violated, we then randomly shuffle the target data $(X_{n+1}', \ldots, X_{n+m}')$ (random permutation in Algorithm 1), leading to a different pair of images in the unsupervised term.

To define a lower bound in this experiment, we simply compute the Dice coefficient across patients employing the provided labels. In Fig. 4, it can be observed that even though the brains across patients are somehow aligned, overlapping between the classes is very low. This is reflected in the poor Dice score reported in Table 3, where the mean Dice value across patients is equal to 42.6. Performing the adaptation with AdapSegNet improves the segmentation performance in the target by roughly 27%. On the other hand, the proposed approach achieves similar results than the adversarial method, while having a reduced complexity. It is worth mentioning that AdapSegNet does not leverage the alignment between images, as shown for example in the T2 to T1 adaptation scenario on iSEG data. When images in $\mathcal{U}$ are aligned, AdapSegNet achieves a mean DICE of 72.61 (Table 2), similar to the case of non-aligned images, i.e., 72.95 (Table 3).

On the contrary, the proposed framework exploits this information, which results in a performance increase of 10.5% and 15.2% when adapting from T2 to T1 and T1 to T2, respectively (Tables 2 and 3), if the data is aligned.
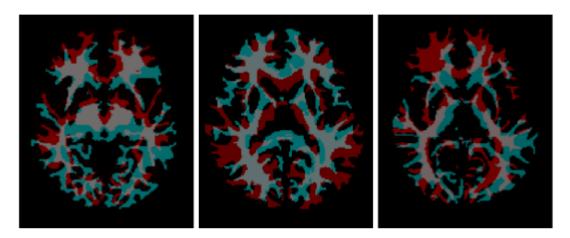

Figure 4: Disalignment found across iSEG dataset subjects. For simplicity, we display only one target class on these images. Different colors represent the WM labels for different patients. Each image shows a different pair of images from different scans. Best viewed in colour.

3.2.3. Training stability.

Adversarial learning strategies are known to be unstable during training. Particularly, for UDA adversarial based approaches this often results in segmentators that produce meaningless out-

			Mean Dice		
			Original alignment Source⟶Target	AdaptSegNet [38] Source⟶Target	Proposed Source⟶Target
Source	Target				
iSEG(T1)	iSEG(T2)	GM	50.4 ± 4.4	67.24 ± 0.96	61.07 ± 2.40
		WM	48.5 ± 4.4	58.65 ± 2.31	51.45 ± 5.24
		CSF	29.0 ± 6.4	75.68 ± 0.43	66.29 ± 2.68
		Mean	42.6 ± 4.7	67.19 ± 1.19	59.60 ± 3.20
iSEG(T2)	iSEG(T1)	GM	50.4 ± 4.4	73.00 ± 0.96	70.86 ± 0.67
		WM	48.5 ± 4.4	64.60 ± 2.73	58.53 ± 1.81
		CSF	29.0 ± 6.4	81.26 ± 2.26	73.36 ± 3.16
		Mean	42.6 ± 4.7	72.95 ± 0.38	67.59 ± 1.46

Table 3: Domain adaptation results on the iSEG dataset when there is misalignment between the images.

puts. Thus, in addition to segmentation performance, we want to compare our method to adversarial in terms of learning convergence. Figure 5 depicts the testing evolution of the mean 3D DICE coefficient for AdaptSegNet and our proposed approach, evaluated every 5 epochs. It can be observed that in both datasets, MRBrains and iSEG, training is very unstable for the adversarial approach, not clearly converging in the MRBrains dataset (*left plot*). As a consequence, the performance can be highly different depending on the number of trained epochs.

On the other hand, the proposed method shows a significantly better stability, smoothly converging as the network is trained.
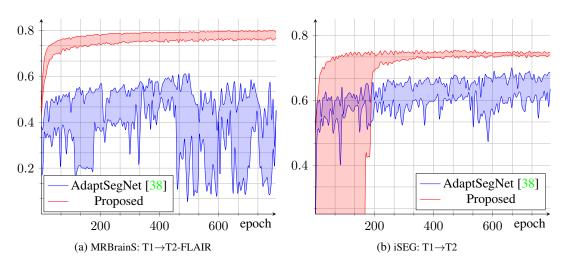

(a) MRBrainS: T1→T2-FLAIR (b) iSEG: T1→T2

Figure 5: Evolution of mean (validation) DICE coefficient over epochs. The minimum and maximum observed value over the three cross-validation runs is plotted, and the area in between is shaded.

12

3.2.4. Impact of the kernel choice

In addition to L2, we also conducted experiments with the Bhattacharyya kernel on both iSEG and MRBrains13 data. As shown in Table 4, we found that the kernel choice has negligible impact on the over-all performance.

		Sq. Euclidian	Bhattacharyya
Source: iSEG(T1) Target: iSEG(T2)	GM	74.12 ± 0.51	74.44 ± 0.43
	WM	66.49 ± 1.08	66.37 ± 0.13
	CSF	83.70 ± 1.04	83.83 ± 0.96
	Mean	74.77 ± 0.61	74.88 ± 0.28
Source: MRB(T1) Target: MRB(T2)	GM	78.25 ± 2.57	79.89 ± 2.47
	WM	82.74 ± 0.89	83.88 ± 0.84
	CSF	73.41 ± 1.25	73.97 ± 1.13
	Mean	78.13 ± 1.36	79.25 ± 1.35

Table 4: Dice coefficients when training with squared Euclidean and Bhattacharyya distances.

4. Conclusions

In this paper, we proposed a simple direct distribution matching approach for unsupervised domain adaptation in the context of semantic segmentation of medical images. Unlike all adversarial approaches in this domain, our method aligns data distributions from both domains in the label space. This output space conveys much richer local and global information, which results in better adapted models. Furthermore, as demonstrated in the experimental results, directly matching output distributions has several benefits compared to adversarial learning: superior performance and better training stability.

Acknowledgments

A. Evolution of DICE coefficients

Figure 6 shows the evolution of the individual DICE coefficients, for gray matter (GM), white matter (WM) and cerebrospinal fluid (CSF), as well as their mean dice, extending Fig. 5. Where available, we also included the proposed algorithm with the Battacharyya kernel.

B. Impact of the kernel choice

In addition to squared Euclidean distance and the Bhattacharyya kernel, we also performed experiments with KL divergence and, again, found that it has negligible impact on the over-all performance. The results, extending Table 4, are detailed in Table 5.

		Sq. Euclidian	Bhattacharyya	KL divergence
	GM	74.12 ± 0.51	74.44 ± 0.43	75.19 ± 0.62
Source: iSEG(T1)	WM	66.49 ± 1.08	66.37 ± 0.13	66.11 ± 0.81
Target: iSEG(T2)	CSF	83.70 ± 1.04	83.83 ± 0.96	83.90 ± 1.19
	Mean	74.77 ± 0.61	74.88 ± 0.28	75.07 ± 0.26
	GM	78.25 ± 2.57	79.89 ± 2.47	79.61 ± 2.38
Source: MRB(T1)	WM	82.74 ± 0.89	83.88 ± 0.84	83.20 ± 0.63
Target: MRB(T2)	CSF	73.41 ± 1.25	73.97 ± 1.13	75.18 ± 0.18
	Mean	78.13 ± 1.36	79.25 ± 1.35	79.33 ± 0.96

Table 5: DICE coefficients when training with squared Euclidean distance, Bhattacharyya distance and KL divergence.

References

[1] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3722–3731, 2017. 2

[2] J. Cai, Z. Zhang, L. Cui, Y. Zheng, and L. Yang. Towards cross-modal organ translation and segmentation: A cycle-and shape-consistent generative adversarial network. *Medical image analysis*, 52:174–184, 2019. 2

[3] C. Chen, Q. Dou, H. Chen, and P.-A. Heng. Semantic-aware generative adversarial nets for unsupervised domain adaptation in chest x-ray segmentation. In *International Workshop on Machine Learning in Medical Imaging*, pages 143–151. Springer, 2018. 2

[4] C. Chen, Q. Dou, H. Chen, J. Qin, and P.-A. Heng. Synergistic image and feature adaptation: Towards cross-modality domain adaptation for medical image segmentation. *arXiv preprint arXiv:1901.08211*, 2019. 3

[5] Y. Chen, W. Li, and L. Van Gool. Road: Reality oriented adaptation for semantic segmentation of urban scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7892–7901, 2018. 2

[6] Y.-H. Chen, W.-Y. Chen, Y.-T. Chen, B.-C. Tsai, Y.-C. Frank Wang, and M. Sun. No more discrimination: Cross city adaptation of road scene segmenters. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1992–2001, 2017. 2

[7] J. Dolz, C. Desrosiers, and I. Ben Ayed. 3D fully convolutional networks for subcortical segmentation in MRI: A large-scale study. *NeuroImage*, 170:456–470, 2018. 2

[8] J. Dolz, K. Gopinath, J. Yuan, H. Lombaert, C. Desrosiers, and I. B. Ayed. Hyperdense-net: A hyper-densely connected CNN for multi-modal image segmentation. *IEEE transactions on medical imaging*, 2018. 2

[9] Q. Dou, C. Ouyang, C. Chen, H. Chen, B. Glocker, X. Zhuang, and P.-A. Heng. PnP-AdaNet: Plug-and-play adversarial domain adaptation network with a benchmark at cross-modality cardiac segmentation. *arXiv preprint arXiv:1812.07907*, 2018. 2, 3

[10] Q. Dou, C. Ouyang, C. Chen, H. Chen, and P.-A. Heng. Unsupervised cross-modality domain adaptation of convnets for biomedical image segmentations with adversarial loss. *arXiv preprint arXiv:1804.10916*, 2018. 2

[11] Y. Ganin and V. Lempitsky. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*, pages 1180–1189, 2015. 2

[12] M. Ghifary, W. B. Kleijn, M. Zhang, D. Balduzzi, and W. Li. Deep reconstruction-classification networks for unsupervised domain adaptation. In *European Conference on Computer Vision*, pages 597–613. Springer, 2016. 2

[13] A. Gholami, S. Subramanian, V. Shenoy, N. Himthani, X. Yue, S. Zhao, P. Jin, G. Biros, and K. Keutzer. A novel domain adaptation framework for medical image segmentation. In *International MICCAI Brainlesion Workshop*, pages 289–298. Springer, 2018. 3

[14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 2

[15] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. A. Efros, and T. Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International Conference on Machine Learning*, pages 1989–1998, 2018. 2

[16] J. Hoffman, D. Wang, F. Yu, and T. Darrell. FCNs in the wild: Pixel-level adversarial and constraint-based adaptation. *arXiv preprint arXiv:1612.02649*, 2016. 2

[17] W. Hong, Z. Wang, M. Yang, and J. Yuan. Conditional generative adversarial network for structured domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1335–1344, 2018. 2

[18] Y. Huo, Z. Xu, H. Moon, S. Bao, A. Assad, T. K. Moyo, M. R. Savona, R. G. Abramson, and B. A. Landman. Synseg-net: Synthetic segmentation without target modality ground truth. *IEEE transactions on medical imaging*, 2018. 2

[19] M. Javanmardi and T. Tasdizen. Domain adaptation for biomedical image segmentation using adversarial training. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 554–558. IEEE, 2018. 3

[20] T. Jebara, R. Kondor, and A. Howard. Probability product kernels. *Journal of Machine Learning Research*, 5:819–844, 2004. 5

[21] K. Kamnitsas, C. Baumgartner, C. Ledig, V. Newcombe, J. Simpson, A. Kane, D. Menon, A. Nori, A. Criminisi, D. Rueckert, et al. Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. In *International Conference on Information Processing in Medical Imaging*, pages 597–609. Springer, 2017. 2, 3

[22] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. 8

[23] S. Klein, M. Staring, K. Murphy, M. A. Viergever, and J. P. Pluim. Elastix: a toolbox for intensity-based medical image registration. *IEEE transactions on medical imaging*, 29(1):196–205, 2010. 7

[24] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017. 2

[25] Y.-C. Liu, Y.-Y. Yeh, T.-C. Fu, S.-D. Wang, W.-C. Chiu, and Y.-C. Frank Wang. Detach and adapt: Learning cross-domain disentangled deep representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8867–8876, 2018. 2

[26] M. Long, Y. Cao, J. Wang, and M. I. Jordan. Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning*, 2015. 2, 3

[27] M. Long, H. Zhu, J. Wang, and M. I. Jordan. Unsupervised domain adaptation with residual transfer networks. In *Advances in Neural Information Processing Systems*, pages 136–144, 2016. 2

[28] A. M. Mendrik, K. L. Vincken, H. J. Kuijf, M. Breeuwer, W. H. Bouvy, J. De Bresser, A. Alansary, M. De Bruijne, A. Carass, A. El-Baz, et al. Mrbrains challenge: online evaluation framework for brain image segmentation in 3t mri scans. *Computational intelligence and neuroscience*, 2015:1, 2015. 6

[29] Z. Murez, S. Kolouri, D. Kriegman, R. Ramamoorthi, and K. Kim. Image to image translation for domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4500–4509, 2018. 2

[30] F. P. Oliveira and J. M. R. Tavares. Medical image registration: a review. *Computer Methods in Biomechanics and Biomedical Engineering*, 17(2):73–93, 2014. 5

[31] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 8

[32] P. Russo, F. M. Carlucci, T. Tommasi, and B. Caputo. From source to target and back: symmetric bi-directional adaptive GAN. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8099–8108, 2018. 2

[33] K. Saito, K. Watanabe, Y. Ushiku, and T. Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3723–3732, 2018. 2

[34] S. Sankaranarayanan and Y. Balaji. Generate to adapt: Aligning domains using generative adversarial networks. In *Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[35] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb. Learning from simulated and unsupervised images through adversarial training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2107–2116, 2017. 2

[36] R. Shu, H. H. Bui, H. Narui, and S. Ermon. A dirtt-t approach to unsupervised domain adaptation. In *Proc. 6th International Conference on Learning Representations*, 2018. 3

[37] B. Sun and K. Saenko. Deep CORAL: Correlation alignment for deep domain adaptation. In *European Conference on Computer Vision*, pages 443–450. Springer, 2016. 2, 3

[38] Y.-H. Tsai, W.-C. Hung, S. Schulter, K. Sohn, M.-H. Yang, and M. Chandraker. Learning to adapt structured output space for semantic segmentation. In *Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 3, 4, 5, 6, 8, 9, 12

[39] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell. Adversarial discriminative domain adaptation. In *Computer Vision and Pattern Recognition (CVPR)*, volume 1, page 4, 2017. 2

[40] T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. *arXiv preprint arXiv:1811.12833*, 2018. 2

[41] L. Wang, D. Nie, G. Li, E. Puybareau, J. Dolz, Q. Zhang, F. Wang, J. Xia, Z. Wu, J. Chen, K. Thung, T. D. Bui, J. Shin, G. Zeng, G. Zheng, V. S. Fonov, A. Doyle, Y. Xu, P. Moeskops, J. P. W. Pluim, C. Desrosiers, I. Ben Ayed, G. Sanroma, O. M. Benkarim, A. Casamitjana, V. Vilaplana, W. Lin, G. Li, and D. Shen. Benchmark on automatic 6-month-old infant brain segmentation algorithms: The iseg-2017 challenge. *IEEE Transactions on Medical Imaging*, pages 1–1, 2019. 6

[42] Z. Wu, X. Han, Y.-L. Lin, M. Gokhan Uzunbas, T. Goldstein, S. Nam Lim, and L. S. Davis. DCAN: Dual channel-wise alignment networks for unsupervised scene adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 518–534, 2018. 2

[43] H. Yan, Y. Ding, P. Li, Q. Wang, Y. Xu, and W. Zuo. Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2272–2281, 2017. 2, 3

[44] Y. Zhang, P. David, and B. Gong. Curriculum domain adaptation for semantic segmentation of urban scenes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2020–2030, 2017. 3

[45] Y. Zhang, S. Miao, T. Mansi, and R. Liao. Task driven generative modeling for unsupervised domain adaptation: Application to x-ray image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 599–607. Springer, 2018. 2, 3

[46] Z. Zhang, L. Yang, and Y. Zheng. Translating and segmenting multimodal medical volumes with cycle-and shape-consistency generative adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9242–9251, 2018. 2

[47] H. Zhao, H. Li, S. Maurer-Stroh, Y. Guo, Q. Deng, and L. Cheng. Supervised segmentation of un-annotated retinal fundus images by synthesis. *IEEE transactions on medical imaging*, 38(1):46–56, 2019. 2, 3

[48] Y. Zou, Z. Yu, B. Vijaya Kumar, and J. Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 289–305, 2018. 3

(a) MRBrainS: T1→T2-FLAIR; GM

(b) MRBrainS: T1→T2-FLAIR; WM

(c) MRBrainS: T1→T2-FLAIR; CSF

(d) MRBrainS: T1→T2-FLAIR; Mean

(e) MRBrainS: T2-FLAIR→T1; GM

(f) MRBrainS: T2-FLAIR→T1; WM

(g) MRBrainS: T2-FLAIR→T1; CSF

(h) MRBrainS: T2-FLAIR→T1; Mean

(i) iSEG: T1→T2; GM

(j) iSEG: T1→T2; WM

(k) iSEG: T1→T2; CSF

(l) iSEG: T1→T2; Mean

(m) iSEG: T2→T1; GM

(n) iSEG: T2→T1; WM
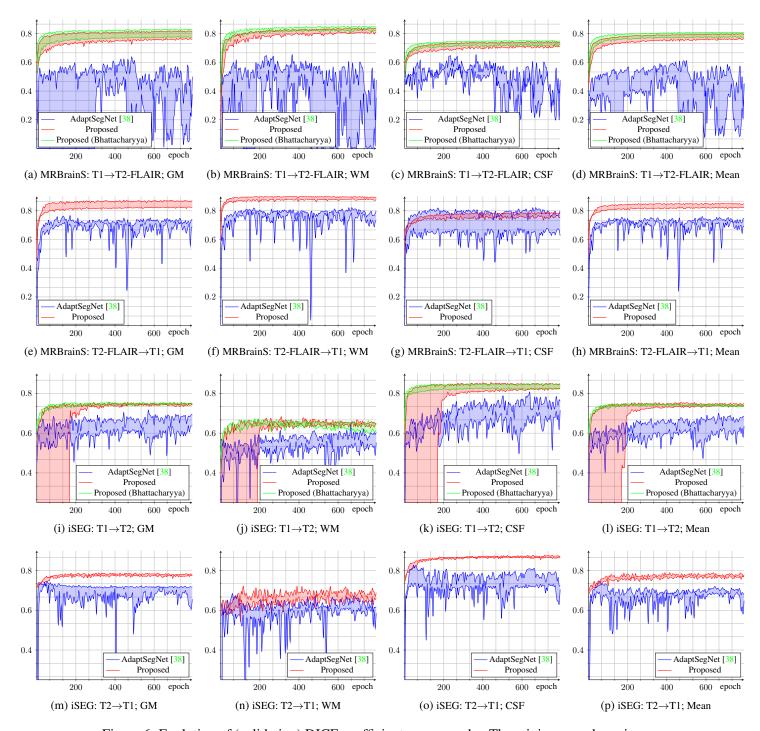
(o) iSEG: T2→T1; CSF

(p) iSEG: T2→T1; Mean

Figure 6: Evolution of (validation) DICE coefficients over epochs. The minimum and maximum observed value over the three cross-validation runs is plotted, and the area in between is shaded.

18