NeuroImage 194 (2019) 1–11

Contents lists available at ScienceDirect

NeuroImage

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

Unsupervised domain adaptation for medical imaging segmentation with self-ensembling

Christian S. Perone [a,*], Pedro Ballester [b], Rodrigo C. Barros [b], Julien Cohen-Adad [a,c]

[a] NeuroPoly Lab, Institute of Biomedical Engineering, Polytechnique Montreal, Montreal, QC, Canada
[b] Machine Intelligence and Robotics Research Group, School of Technology, Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, RS, Brazil
[c] Functional Neuroimaging Unit, CRIUGM, Universite de Montreal, Montreal, QC, Canada

A B S T R A C T

Recent advances in deep learning methods have redefined the state-of-the-art for many medical imaging applications, surpassing previous approaches and sometimes even competing with human judgment in several tasks. Those models, however, when trained to reduce the empirical risk on a single domain, fail to generalize when applied to other domains, a very common scenario in medical imaging due to the variability of images and anatomical structures, even across the same imaging modality. In this work, we extend the method of unsupervised domain adaptation using self-ensembling for the semantic segmentation task and explore multiple facets of the method on a small and realistic publicly-available magnetic resonance (MRI) dataset. Through an extensive evaluation, we show that self-ensembling can indeed improve the generalization of the models even when using a small amount of unlabeled data.

1. Introduction

In the past few years, the research community has witnessed the fast developmental pace of deep learning (LeCun et al., 2015) approaches for unstructured data analysis, arguably establishing an important scientific milestone. Deep neural networks constitute a paradigm shift from traditional machine learning approaches for unstructured data. Whereas the latter rely on hand-crafted feature engineering for improving learning over images, text, audio, and similarly unstructured inputs, deep neural networks are capable of automatically learning robust hierarchical features, in what is known as *representation learning*. Deep learning approaches have achieved human-level performance on many tasks and, indeed, sometimes even surpassing it in applications such as natural image classification (He et al., 2016), or arrhythmia detection (Rajpurkar et al., 2017).

Due to its popularity and compelling results in many domains, deep learning attracted a lot of attention from the medical imaging community. A recent survey by Litjens et al. (2017) analyzed more than 300 medical imaging studies, and found that deep neural networks have become pervasive throughout the field of medical imaging, with a significant increase in the number of publications between 2015 and 2016. The survey also identified that the most addressed task is image segmentation, likely due to the importance of quantification of anatomical structures and pathologies (Gros et al., 2018) for disease diagnosis and prognosis, as opposed to less informative tasks such as classification of pathologies or detection of structures, which can be posed as a segmentation tasks as well, but not the opposite.

Deep neural networks are thus becoming the norm in medical imaging, though there are still several unsolved challenges that remain to be addressed. For instance, one of the most well-known problems is the high sample complexity, or how much data deep learning requires to accurately learn and perform well on unseen images, which is related to the concepts of model complexity and generalization, active areas of research in learning theory (Neyshabur et al., 2017).

The large amount of required data to train deep neural networks can be partially mitigated with techniques such as transfer learning (Yosinski et al., 2014; Zamir et al., 2018). However, transfer learning is problematic in medical imaging because a large dataset is still required so the models can benefit from the inductive transfer process. Unlike the case of natural images, where annotations can be easily provided by non-experts, medical images require careful and time-consuming analysis from trained experts such as radiologists.

Yet another challenge when deploying deep learning models to medical imaging analysis – and perhaps one of the most difficult to solve – is the so-called *data distribution shift*, wherein different imaging scenarios (e.g. parameter choices, different protocols) can result in vastly different data distributions, despite imaging a common object. Therefore, models trained under the empirical risk minimization (ERM) principle, might fail to generalize to other domains due to its strong assumptions. ERM is the statistical learning principle behind many machine learning

* Corresponding author.
E-mail address: christian.perone@gmail.com (C.S. Perone).

https://doi.org/10.1016/j.neuroimage.2019.03.026
Received 10 January 2019; Received in revised form 4 March 2019; Accepted 12 March 2019
Available online 19 March 2019
1053-8119/

methods, and it offers good learning guarantees and bounds if its assumptions hold, such as the fact that the training and test datasets derive from similar domains. However, in practice, this assumption is often violated.

When a deep learning model that assumes independent and identically-distributed (iid) data is trained with images from one domain and is subsequently deployed on images from a different domain (e.g. distinct imaging center), that follow a distinct data distribution, its performance often degrades by a large margin. An example of domain shift can be seen in magnetic resonance imaging (MRI) in different centers, where machine vendor, software versions, radio-frequency coils, and sequence parameters (e.g., slice positioning, image resolution) often vary, producing images that come from different distributions. Fig. 1 illustrates those inter-center differences in data distribution, based on data from the Gray Matter (GM) segmentation challenge (Prados et al., 2017). Fig. 2 illustrates the associated voxel intensity distribution for the same dataset.

Although this distribution shift is common in medical imaging, the problem is surprisingly ignored during the design of many different challenges in the field. It is common to have the same domain data (same machine, protocol, etc.) on both training and test sets. However, this homogeneous data split often does not represent the reality and in many cases will produce over-optimistic evaluation results. In practice, it is rare to have labeled data available from a new center before training a model, hence it is common to use a pre-trained model from a different domain on completely different data. Therefore, it is paramount to have a proper evaluation avoid contaminating the test set with data from the same domain that is present in the training set. Incurring the risk of the detrimental effects of inadequate evaluations (Zech et al., 2018). The name given to learn a classifier model or any other predictor with a shift between the training and the target/test distributions is known as "domain adaptation" (DA). In this work we expand upon a previously-developed method (French et al., 2017) for DA based on the Mean Teacher (Tarvainen and Valpola, 2017) approach, to segmentation tasks, the most addressed task in medical imaging.

We provide the following contributions: (i) we extend the unsupervised DA method using self-ensembling for the semantic segmentation task; to the best of our knowledge, this is the first time this method is used for semantic segmentation in medical imaging; (ii) we explore some model components such as different consistency losses, and evaluate the performance of our method on a series of experiments using a realistic small MRI dataset; (iii) we perform an ablation experiment to provide strong evidence that unlabeled data is responsible for the observed performance improvement, ruling out the effects of the exponential moving average; (iv) we provide visualizations to derive insight into the model dynamics of the unsupervised DA task.

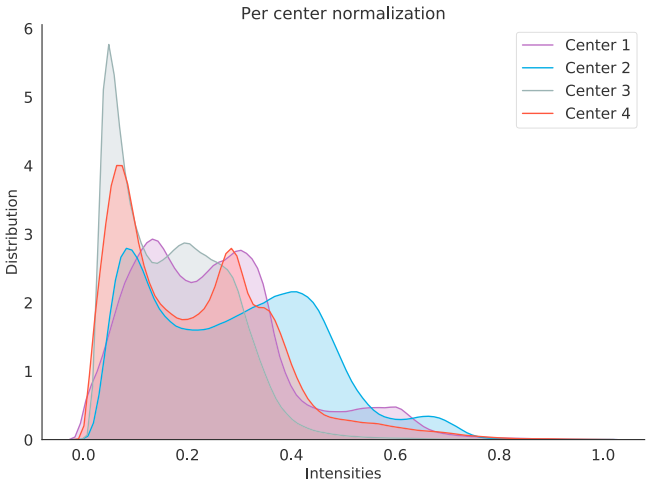

Fig. 2. MRI axial-slice pixel intensity distribution from four different centers (UCL, Montreal, Zurich, Vanderbilt) that collaborated to the SCGM Segmentation Challenge (Prados et al., 2017). Normalized between 0 and 1 per center.

This paper is organized as follows. In Section 2 we present related work, in Section 3 we give a brief treatment to the unsupervised DA task and its connection to semi-supervised learning. In Section 4 we describe our method in terms of model architecture and corresponding design decisions. In Section 5 we describe the dataset used in our experiments and how we performed the data split for the DA scenario. In Section 6 we provide the experiment results, followed by an ablation study in Section 7. In Section 8 we provide visual insights from the adaptation dynamics

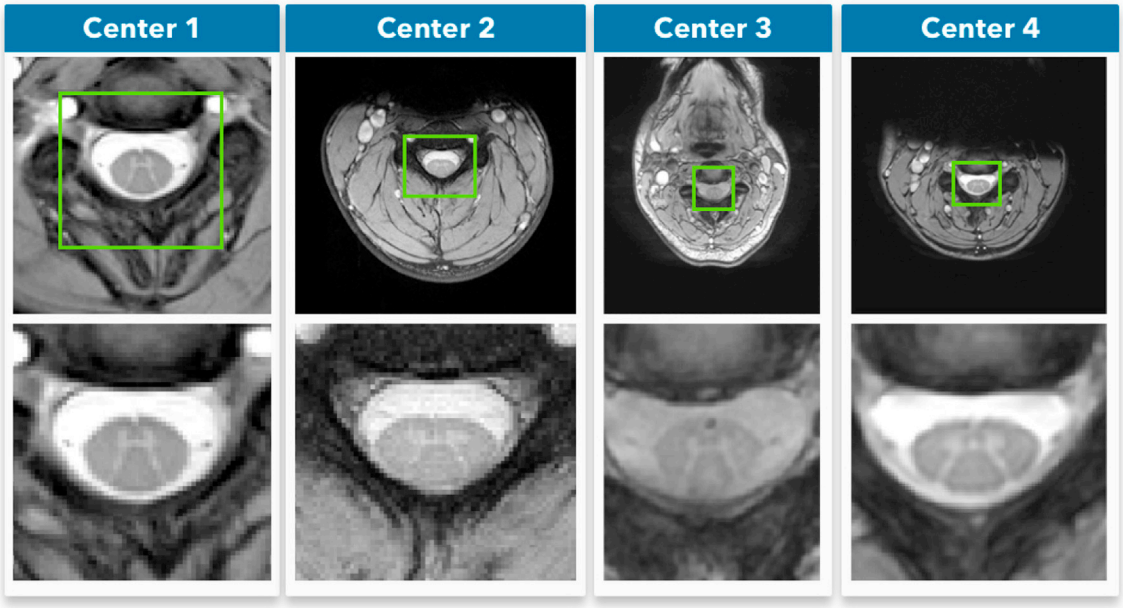

Fig. 1. Samples of axial MRI from four different centers (UCL, Montreal, Zurich, Vanderbilt) that participated in the SCGM Segmentation Challenge (Prados et al., 2017). **Top row**: original MRI images. **Bottom row**: crop of the spinal cord (green rectangle). Reproduced from (Perone and Cohen- Adad, 2018b). Best viewed in color.

of the model for multiple domains. Finally, in Section 9 we discuss our findings and limitations of our work. In the spirit of open science and reproducibility, we also provide more information regarding data and source-code availability in Section 10.

2. Related work

Deep learning methods for medical imaging has become a popular research focus in recent years (Litjens et al., 2017). Before the development of deep learning models, initial work was focused mostly on patch-based (Coupé et al., 2011) segmentation. With the growing interest in deep learning for computer vision, the first attempts using Convolutional Neural Networks (CNNs) for image segmentation processed image patches through a sliding window, to yield segmented patches, which were then stitched together to yield the final segmented image (Lai, 2015). The main drawbacks of this approach are computational cost (i.e., several forward passes are required to produce the segmented images) and inconsistency in predictions, the latter of which can be fixed or partially mitigated by overlapping sliding windows, depending on the network architecture.

Though patch-wise methods continue to be actively researched (Hou et al., 2016) and have led to several advances in segmentation (Lai, 2015), presently, the most common deep architecture for segmentation is or is based on the so-called Fully Convolutional Network (FCN) (Long et al., 2015). This architecture is solely based on convolutional layers with the final result not depending on the use of fully-connected layers. FCNs can provide a fully-segmented image within a single forward step, and with variable output size depending on the size of the input tensor. One of the most well-known FCNs for medical imaging is the U-net (Ronneberger et al., 2015), which combines convolutional, down-sampling, and upsampling operations with skip non-residual connections. In this work we used the U-Net architecture, although the proposed framework is decoupled from the choice of network architecture, as further discussed in Section 4.3.

Deep Domain Adaptation (DDA), which is a field unrelated in essence to medical imaging, has been widely studied in the recent years (Wang & Deng, 2018). We can divide the literature on DDA as follows: (i) methods based on building domain-invariant feature spaces through auto-encoders (Ghifary et al., 2016), adversarial training (Ganin et al., 2016), GANs (Hoffman et al., 2017; Sankaranarayanan et al., 2018), or disentanglement strategies (Liu et al., 2018; Cao et al., 2018); (ii) methods based on the analysis of higher-order statistics (Li et al., 2016; Sun and Saenko, 2016); (iii) methods based on explicit discrepancy between source and target domains (Tzeng et al., 2014); and (iv) self-ensembling methods based on implicit discrepancy (French et al., 2017; Tarvainen and Valpola, 2017).

In (Hoffman et al., 2017), the authors trained GANs with cycle-consistent loss functions (Zhu et al., 2017) to remap the distribution from the source to the target dataset, thereby creating target domain specific features for completing the task. In (Sankaranarayanan et al., 2018), GANs were employed as a means of learning aligned embeddings for both domains. Similarly, disentangled representations for each domain have been proposed (Liu et al., 2018; Cao et al., 2018) with the goal of generating a feature space capable of separating domain-dependent and domain-invariant information.

In (Li et al., 2016), the authors proposed to change parameters of the neural network layers for adapting domains by directly computing or optimizing higher-order statistics. More specifically, they proposed an alternative for batch normalization called Adaptive Batch Normalization (AdaBN) that computes different statistics for the source and target domains, hence creating domain-invariant features that are normalized according to the respective domain. In a similar fashion, Deep CORAL (Sun and Saenko, 2016) provides a loss function for minimizing the covariances between target and source domain features.

Discrepancy-based methods pose a different approach to DDA. By directly minimizing the discrepancy between activations from the source and target domain, the network learns to generate reasonable predictions while incorporating information from the target domain. The seminal work of Tzeng et al. (2014) directly minimizes the discrepancy between a specific layer with labeled samples from the source set and unlabeled samples from the target set.

Implicit discrepancy-based methods such as self-ensembling (French et al., 2017) have become widely used for unsupervised domain adaptation. Self-ensembling is based on the Mean Teacher network (Tarvainen and Valpola, 2017), which was first introduced for semi-supervised learning tasks. Due to the similarity between unsupervised domain adaptation and semi-supervised learning, there are very few adjustments that need to be made to employ the method for the purposes of DDA. Mean Teacher optimizes a task loss and a consistency loss, the latter minimizing the discrepancy between predictions on the source and target dataset. We further detail how Mean Teacher works in Section 4.1.

There are a few studies that report results of using different data domains for medical imaging by making use of the unsupervised domain adaptation literature. The work (AlBadawy et al., 2018) discusses the impact of deep learning models across different institutions, showing a statistically significant performance decrease in cross-institutional train-and-test protocols. A few studies have applied domain adaptation to medical imaging directly by using adversarial training (Kamnitsas et al., 2017; Chen et al., 2018; Zhang et al., 2018; Lafarge et al., 2017; Javanmardi and Tasdizen, 2018; Dou et al., 2018), with some studies using generative models to augment training (Mahmood et al., 2018; Madani et al., 2018). Nevertheless, to the best of our knowledge, this present work is the first to address the problem of domain shift in medical image segmentation by extending the unsupervised DA self-ensembling method to semantic segmentation tasks.

3. Semi-supervised learning and unsupervised domain adaptation

A common approach for improving training when few labeled examples are available is semi-supervised learning, which is defined as follows: given a labeled dataset with distribution $P(X_l)$ and unlabeled data with distribution $P(X_u)$, learn from both labeled and unlabeled data in order to improve a supervised learning task (say, classification) or an unsupervised learning task (say, clustering).

Semi-supervised learning methods tend to perform well when unlabeled data actually come from the same distribution as the labeled data. This allows the learning algorithm to leverage its knowledge using unlabeled data, which usually represents the majority of samples. As promising as semi-supervised learning is, the assumption that the distribution of unlabeled data $P(X_u)$ is similar to $P(X_l)$ often fails in real-world applications. We refer the reader to a thorough evaluation of semi-supervised learning methods and their limitations in (Odena et al., 2018).

It often happens that models are applied in situations that are largely different from those in which they were originally trained. Examples include different weather conditions for outdoor activity recognition, or different cities for training driverless vehicles. Those changes in scenario shift the data distribution $P(X)$, reducing the quality of the predictions in cases where the model was not properly adapted to the desired condition.

The difference between the distributions from the examples used in training and test sets is called *domain shift*. Consider a source dataset with input distribution $P(X_s)$ and label distribution $P(Y|X_s)$, as well as a target dataset with input distribution $P(X_t)$ and labels $P(Y|X_t)$, $P(X_s) \neq P(X_t)$. Domain adaptation can be addressed via a supervised approach where labeled data from the target domain is available, or via unsupervised learning where only unlabeled data is available for the target domain.

When a method addresses the problem of domain adaptation using unlabeled data for the target domain, which is the most common and useful scenario, the task at hand is called *unsupervised domain adaptation*. Unsupervised domain adaptation methods assume that distributions $P(X_s)$, $P(Y|X_s)$ and $P(X_t)$ are available, while $P(Y|X_t)$ is not. In other words, only the source dataset provides labeled examples. Hence, the

task is to leverage knowledge from the target domain using the unlabeled data available in $P(X_t)$.

4. Method

This section details the base domain adaptation methods that we used for the medical image application. We further discuss the changes that are needed to enable unsupervised domain adaptation for segmentation tasks, as opposed to the typical classification scenario.

4.1. Self-ensembling and mean teacher

Self-ensembling was originally conceived as a viable strategy for generating predictions on unlabeled data (Laine and Aila, 2016). The original paper proposes two different models for self-ensembling. The first model, called Π, employs a consistency loss between predictions on the same input. Each input from a batch is passed twice through a neural network, each time with distinct augmentation parameters, to yield two different predictions. A squared difference between those predictions is minimized along with the cross-entropy for the labeled examples. The second model, called temporal ensembling, works under the assumption that as the training progresses, averaging the predictions over time on unlabeled samples may contribute to a better approximation of the true labels. This pseudo-label is then considered as a target during training. The squared difference between the averaged predictions and the current one is minimized along with the cross-entropy for labeled examples. The network performs the exponential moving average (EMA) to update the generated targets at every epoch:

$$f'(x)_t = \alpha f'(x)_{t-1} + (1-\alpha) f(x)_t \tag{1}$$

Where t is the step, x is the data, $f(\cdot)$ is the network and α is a momentum term that controls how far the ensemble reaches training history data.

Self-ensembling was extended to directly combine model weights instead of predictions. This adaptation is called the Mean Teacher (Tarvainen and Valpola, 2017) model. Considering Eq. (1) for updating the target pseudo-labels, Mean Teacher updates the model weights at each step, thus generating a slightly improved model compared to the model without the EMA, a framework which is linked to the Polyak-Ruppert Averaging (Polyak and Juditsky, 1992; Ruppert, 1988). In this scenario, the EMA model was named teacher, and the standard model, student. The update function is as follows:

$$\theta'_t = \alpha \theta'_{t-1} + (1-\alpha)\theta_t \tag{2}$$

where θ are the model parameters, t is the step and α is the hyperparameter regulating the importance of the current model's weights with respect to previous models. The best results are achieved when α is increased later on during training, as the student is close to convergence, causing the teacher model to benefit from having a larger memory of its past weights.

Each training step involves a loss component for both labeled and unlabeled data. All samples from a batch are evaluated by both the student and teacher models, with their respective predictions compared via the consistency loss. The labeled data, however, is also compared to its ground truth, as traditionally performed in segmentation tasks, in what we call the task loss:

$$J(\theta) = J_{task}(\theta) + \gamma J_{consistency}(\theta) + \lambda R(\theta) \tag{3}$$

where γ and λ are the Lagrange multipliers that represent, respectively, the consistency and regularization weights. The γ hyperparameter was empirically found to improve results when it varied through time, given that in the earlier training steps the network continues to generate poor results. The consistency weight follows a sigmoid ramp-up saturating at a given user-defined value.

Mean Teacher follows the dynamics of model distillation (Hinton et al., 2015). In this scenario, a trained model is used for predicting instances and its output is used as labels for another, smaller model. This is considered a good practice as soft labels tend to better represent the characteristics of the classes (e.g., the representation distance between a Siberian Husky and an Alaskan Malamute should arguably be smaller than the distance between a Siberian Husky and a Persian Cat). Unlike traditional distillation formulations, the Mean Teacher framework also uses the teacher model to generate labels for unlabeled data and represents a model of the same size that is simultaneously updated during training.

The Mean Teacher framework was also extended for unsupervised domain adaptation in (French et al., 2017). Among the proposed changes, the authors modified the data batches such that each batch consists of images from both the source and target domains. At each step, the student model evaluates images from the source domain and computes derivatives via a task loss based on the ground truth. The target domain images, which are unlabeled, are used to compute the consistency loss by comparing predictions from both student and teacher models. It differs from its original formulation in that the teacher model only has access to unlabeled examples (in this case, examples from the target domain). Each loss function is thus responsible for improving learning at a single domain. The task loss is evaluated by comparing the predictions against the ground truth for the labeled examples (source domain). For the consistency loss, MSE is often used to evaluate the predictions from both student and teacher models for the unlabeled examples (target domain).

4.2. Adapting mean teacher for segmentation tasks

Both the original and adapted Mean Teacher versions for unsupervised domain adaptation rely on the cross-entropy classification cost. Given that we are not dealing with classification but with a segmentation task, we need to minimize a different loss function that takes into consideration the particularities of that task. Originally proposed in (Milletari et al., 2016), the Dice loss generates reliable segmentation predictions due to its low sensitivity to class imbalance:

$$J_{task}(\theta) = -\frac{2^* \sum_i^N p_i g_i}{\sum_i^N p_i + \sum_i^N g_i} \tag{4}$$

where p_i and g_i are flattened predictions and ground truth values for an instance, respectively. Dice was kept as the task loss for both baseline and adaptation experiments. Note that the dice loss is computed for the entire batch at once, unlike the typical strategy of averaging when using cross-entropy, for instance.

A second problem when training the student and teacher models for segmentation tasks is the inconsistency introduced between training samples of the student and teacher models when a spatial transformation (e.g., translation, rotation, or any similar spatial transformation for the purpose of data augmentation) is applied with different parameters to both inputs of the teacher and student models.

To solve that problem we used the same approach employed by (Perone and Cohen-Adad, 2018a) as shown in Fig. 4. The spatial transformation $g(x;\phi)$, where x is the input data and ϕ are the transformation parameters (i.e., rotation angle), is applied to the student model before feeding data into the model. For the teacher model, the same transformation $g(x;\phi)$ is applied to the predictions of the teacher model, causing both predictions to be aligned for the consistency loss. This framework is possible because backpropagation only occurs for the student model and therefore there is no need for differentiation on the delayed augmentation of the teacher model. The proposed method is illustrated in Fig. 4. Examples of images after data augmentation and their respective compensated ground truth are shown in Fig. 3.

We decided to only conduct data augmentation at the slice-level, not taking into consideration column-wise cord deformations. This makes

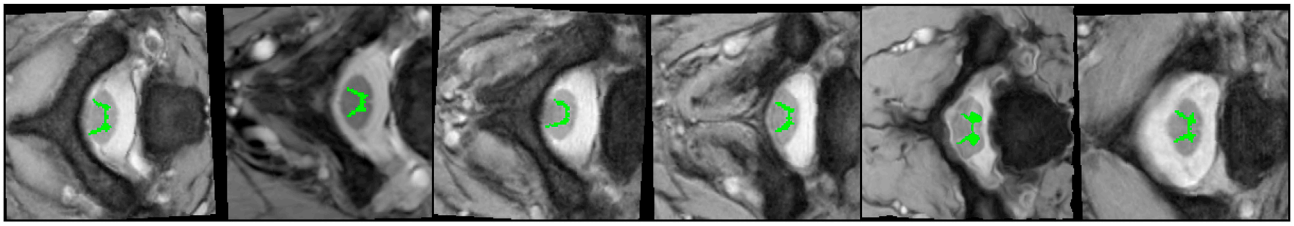

Fig. 3. Random and data-augmented MRI axial-slice samples from the SCGM Segmentation Challenge (Prados et al., 2017). We show how the ground-truth (shown in green) has to be rotated with the same parameters as the slice in order to match the desired region.

implementation easier and also more generalizable for datasets that do not have access to full-volume data, with the cost of maybe not producing realistic results. Another advantage is the possibility to generate augmented data during training without much computational cost. The actual augmentation protocol we follow is an elastic transformation, a random affine transform, and a random tensor channel shift that adds or subtract a value which is sampled from a uniform distribution into the entire channel (voxel-wise).

4.3. Model architecture

Since the U-net (Ronneberger et al., 2015) is widely applied in medical imaging field for diverse tasks, in order to provide results that can generalize to a wide spectrum of applications, for all experiments we employed the U-net (Ronneberger et al., 2015) model architecture with 15 layers, group normalization (Wu and He, 2018), and dropout. The rationale behind group normalization and not batch normalization is discussed later.

To provide a fair comparison, we followed the recommendations from (Oliver et al., 2018) and kept the same model for the baseline and for our method. While the Mean Teacher model also acts as a regularizer, we kept the same regularization weights for all comparisons. Regularization weights can be fine-tuned, however, possibly improving even further the results of Mean Teacher.

4.4. Baseline employed

Our baseline is exactly the same U-Net architecture that was employed for the teacher and student models in our method. The training hyper-parameters are also the same as used on our method, in order to make a fair comparison among the baseline and our proposal, following the same methodology used in (Oliver et al., 2018). The only difference between the baseline and our method is that whereas in the baseline we train the model in a standard supervised learning fashion with no additional unlabeled data, in our method we follow the framework protocol described in Section 4.

We conducted an extensive hyperparameter search to find a proper baseline model, yielding a mini-batch size of 12 and a dropout rate of 0.5. For training, we used the Adam optimizer (Kingma and Ba, 2015) with L_2 penalty factor of $\lambda = 6 \times 10^{-4}$, $\beta_1 = 0.99$, and $\beta_2 = 0.999$. For learning rate, we used a sigmoid learning rate ramp-up strategy until epoch 50 followed by a cosine ramp-down until epoch 350. Eq. (5) shows the sigmoid ramp-up strategy:

$$R_{up}(m) = \alpha e^{-5(1-m)^2} \tag{5}$$

where α is the highest learning rate and m represents the ratio between current epoch and the total ramp-up epochs. Eq. (6) presents the cosine ramp-down strategy:

$$R_{down}(r) = \alpha \frac{\cos(\pi r) + 1}{2} \tag{6}$$

where α is the highest learning rate and r is the ratio between the number of epochs after the ramp-up procedure and the total number of epochs expected for training.

For a fair comparison, and to be able to assess the specific benefits of domain adaptation, no hyperparameter from the baseline model was changed in the adaptation scenario. The only change concerned the hyperparameters, which only affect the domain adaptation training procedure.

4.5. Consistency loss

The consistency loss is one of the most important aspects of Mean Teacher. If the difference between predictions from teacher and student is not representative enough for distilling the knowledge on the student model, the method will not properly work or training may even diverge. In the original implementation of the Mean Teacher method, the mean squared error (MSE) was proposed:

$$J_{MSE}(\theta) = \frac{\sum_i^N (p_i - g_i)^2}{N} \tag{7}$$

where p_i and g_i are flattened predictions from student and teacher, respectively.

As an alternative, the cross-entropy is more commonly used for classification tasks. The cross-entropy is defined as:

$$J_{CE}(\theta) = -\sum_i^N p_i \log g_i \tag{8}$$

where p_i and g_i are predictions from student and teacher, respectively. However, cross-entropy is also known to be sensitive to class imbalance.

Our preliminary experiments led to use MSE with different weights per class to address the problem of class imbalance. However, this approach relies on thresholding predictions from the teacher to define binary expected voxel values for the student. Defining both the correct weights and the threshold value is a difficult task that did not seem to improve overall results.

The same problem happens with more complex losses, e.g., the Focal Loss (Lin et al., 2018), due to additional hyperparameters (in this case, γ and β).

We have thus explored other losses: the Dice loss, presented in Section 4, and the Tversky loss (Salehi et al., 2017). The Tversky loss is a variation of the dice loss that aims at mitigating the problem of class imbalance, which is common in medical image segmentation tasks. It is defined as:

$$J_{tversky}(\theta) = -\frac{\sum_i^N p_{0i} g_{0i}}{\sum_i^N p_{0i} g_{0i} + \alpha \sum_i^N p_{0i} g_{1i} + \beta \sum_i^N p_{1i} g_{0i}} \tag{9}$$

where p_{0i} and g_{0i} represent the predicted probabilities and expected ground-truth of a voxel that belongs to the correct tissue, whereas p_{1i} and g_{1i} respectively represent the predicted probabilities and expected ground-truth (0 or 1) of a voxel that belongs to any other tissue. The α

and β hyperparameters address the problem of class imbalance. The Tversky loss, however, is hampered by the difficulty of determining more hyperparameters alongside the consistency weight value (same issue as noted above with the weighted MSE).

We have also noticed that both Dice and Tversky coefficients are problematic when used as consistency losses. Albeit properly representing the nature of the task, their formulation is based on multiplication and it is assumed that the ground-truth is binary, i.e. $g_i \in \{0, 1\}$. However, given that we use the teacher soft outputs (i.e., not binary), both Dice and Tversky losses do not obey the proper score orientation: $S(G,y) > S(G^*,y)$, where S is the scoring function and y is the ground truth. This relationship should hold only if G is a better probabilistic forecast, which is not the case for Tversky and Dice when using soft targets.

For example, if $p_i = 0.9$ and $g_i = 1.0$, the numerator yields 0.9. However, when $p_i = 0.9$ and $g_i = 0.9$, the score should increase (because the predicted and ground-truth are the same), but instead the numerator decreases to 0.81 and the output score also decreases.

One way to overcome this issue is to threshold the teacher's predictions such that the loss functions can work as expected. However, identifying suitable threshold values is not trivial since they drastically impact how the network adapts, and reduces the benefits of using a distillation-based (Hinton et al., 2015) approach. An alternative to thresholding is to modify the formulations of the loss functions such that they can properly handle non-binary labels. A detailed analysis of such modifications falls outside the scope of this paper so we left it for future work.

4.6. Batch normalization and group normalization for domain adaptation

Batch Normalization (Ioffe and Szegedy, 2015) (BN) is a method used to improve the training of deep neural networks through the stabilization of the distribution of layer inputs. Nowadays, Batch Normalization is pervasive in most deep learning architectures, enabling the use of large learning rates and helping with convergence.

Initially thought to help with the internal covariate shift (ICS) problem (Ioffe and Szegedy, 2015), Batch Normalization was recently found (Santurkar et al., 2018) to smooth the optimization landscape of the network due to the improvement of the Lipschitzness, or β-smoothness (Santurkar et al., 2018) of both loss and gradients.

Batch Normalization works differently for training and inference. During training, the normalization happens using the batch statistics, while on inference it uses the population statistics, usually estimated with moving averages on each batch during the training procedure. This strategy, however, is problematic for domain adaptation via Mean Teacher, given that there are multiple distributions being fed during training, causing the Batch Normalization statistics to be computed with both source and target data.

One possible approach to overcome that issue is to use different batch statistics for the source and the target domains as done in AdaBN (Li et al., 2016). Implementing this approach within the training procedure is easily achieved using modern frameworks because it only requires to forward the batch to each domain separately (French et al., 2017). However, in the implementation of French et al., both source and target domain data were used to compute the running average at inference. One should ideally perform running averages and population statistics on both domains separately, though at the expense of increased complexity on training, especially when running on a multi-GPU scenario with small batch sizes, a very common scenario in segmentation tasks where synchronization is also required. Another alternative to Batch Normalization limitations is to use Weight Normalization (Salimans and Kingma, 2016) instead. In Weight Normalization, the weight vectors have their norm controlled in order to improve convergence. Although not using mini-batch statistics — which is our main concern on BN — research showed that it fails to compete with BN in many tasks (Wu and He, 2018).

Besides the mentioned issues, Batch Normalization also suffers from sub-optimal results when using small batch sizes (Wu and He, 2018), which are very common in segmentation tasks due to memory requirements. For those reasons, we chose Group Normalization (Wu and He, 2018), an alternative to Batch Normalization where channels are divided into groups and where mean and variance are computed within each group regardless of batch sizes. Group Normalization works consistently better than Batch Normalization with small batch sizes (typically ¡15) and does not require storing running averages for the population statistics, simplifying the training and inference procedures and providing better results for our scenario that involves domain adaptation and segmentation tasks.

4.7. Hyperparameters for unsupervised domain adaptation

A problem shared by many techniques for unsupervised domain adaptation is how to set proper hyperparameters such as the learning rate or the consistency weight. In unsupervised settings, there are no labeled data from the target domain so the estimation of hyperparameters from the source distribution alone can be completely different from those from the target distribution.

An alternative method to solve this issue is to use reverse cross-validation (Zhong et al., 2010), which was also used in (Ganin et al., 2016). The variant of this method, as used in (Ganin et al., 2016) works as following: given the labeled source sample S and the unlabeled target sample T, each set is split into training sets and validation sets (S_V and T_V respectively) (Ganin et al., 2016).

The labeled set S' and the unlabeled target set T' are then used to learn a classifier η. Using the same algorithm, a reverse classifier η_r is learnt using the self-labeled set $\{(x, \eta(x))\}_{x \in T'}$ and the unlabeled part of S' as target sample. The reverse classifier η_r is then evaluated on the validation set S_V of source sample (Ganin et al., 2016).

However, once again, this approach comes at the expense of increasing the complexity of the validation process. Nevertheless, we found that the estimation of hyperparameters for Mean Teacher on the source domain yielded robust results, therefore we adopted them in our experiments. We are aware that such a simple strategy is a limitation of our evaluation procedure since we could probably achieve better results for our proposed method by incorporating a more sophisticated hyperparameter estimation procedure.

5. Materials

The Spinal Cord Gray Matter Challenge (Prados et al., 2017) dataset is a multi-center, multi-vendor, and publicly-available MRI data collection that is comprised of 80 healthy subjects with 20 subjects from each center.

The demographics of the dataset range from a mean age of 28.3 up to 44.3 years old. Three different MRI systems were employed (Philips Achieva, Siemens Trio, Siemens Skyra) with distinct acquisition parameters. The voxel size resolution of the dataset ranges from $0.25 \times 0.25 \times 2.5$ mm up to $0.5 \times 0.5 \times 5.0$ mm and the number of axial slices ranged from 3 to 28. The dataset is split between training (40) and test (40) sets, and the test set labels are hidden (not publicly available). For each labeled slice in the dataset, 4 gold-standard segmentation masks were manually created by 4 independent experts (one per participating center). For more detailed information regarding the dataset (e.g., the MRI parameters), please see (Prados et al., 2017). We considered each slice/rater pair an independent sample, thus using 4 times the number of slices for training and testing than the total number of samples.

Since the Spinal Cord Gray Matter Challenge dataset contains data from all 4 centers both in the training and test sets, we used a non-standard split in order to evaluate our technique within the domain adaptation scenario, where the domain present in the test set is not contaminated by the training data domain. Therefore, we used centers 1 and 2 as the training set, center 3 as the validation set, and center 4 as the test set.

We used the unlabeled data from center 4 test set (which does not

contain publicly-available labels) as the unlabeled data for the target domain, and we used the training data from center 4 (with labels) as the test set to evaluate the final performance of our model. We also slice all 3D samples into 2D axial slices and resampled each slice to 0.25×0.25 mm. An overview of the dataset is presented in Fig. 5.

6. Experiments

We have designed several experiments to understand the behavior of different aspects of domain adaptation on the medical imaging domain. We have also performed ablation studies and evaluated multiple metrics for each center.

6.1. Adapting to different centers

We trained the network with both centers 1 and 2 in a supervised fashion. We then adapted the network to centers 3 and 4 separately. With this setup, we were able to address three related research questions on adaptation and semi-supervised learning:

1. How do predictions change at inference time when images from domains different than the source domain are presented?
2. How does the network change its predictions to the novel domain after performing domain adaptation?
3. How well does an adapted network generalize when presented with images that were not used during training, neither as a supervised signal nor as an unsupervised adaptation component?

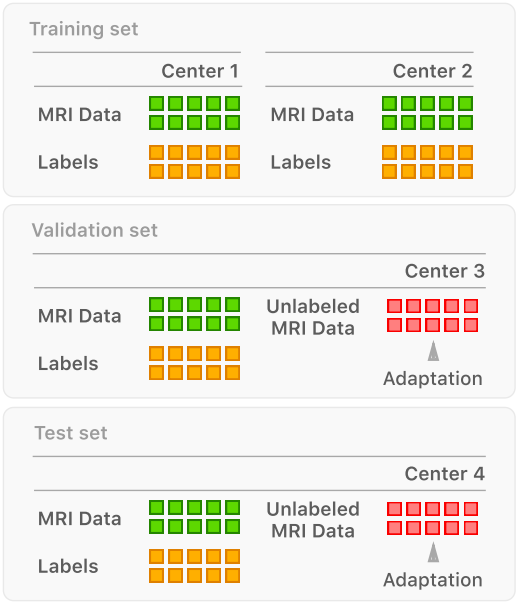

Fig. 5. Overview of the data splitting method for training machine learning models. Each colored square represents a single subject of the dataset (containing multiple axial slices).

Results of this first experiment are presented in Table 1, where all metrics were computed on axial slices on a 2D fashion and in the resampled target space.

Regarding *Question 1*. Both centers 1 and 2 are included in the training set and we would like to assess whether additional unsupervised data from different domains (centers 3 or 4) improve generalization on the centers 1 and 2. For both adapted centers 3 and 4, results for all metrics (except for recall) outperform the baseline, suggesting a positive change in prediction performance for the source domain after domain adaptation on unseen domains leveraging unlabeled data.

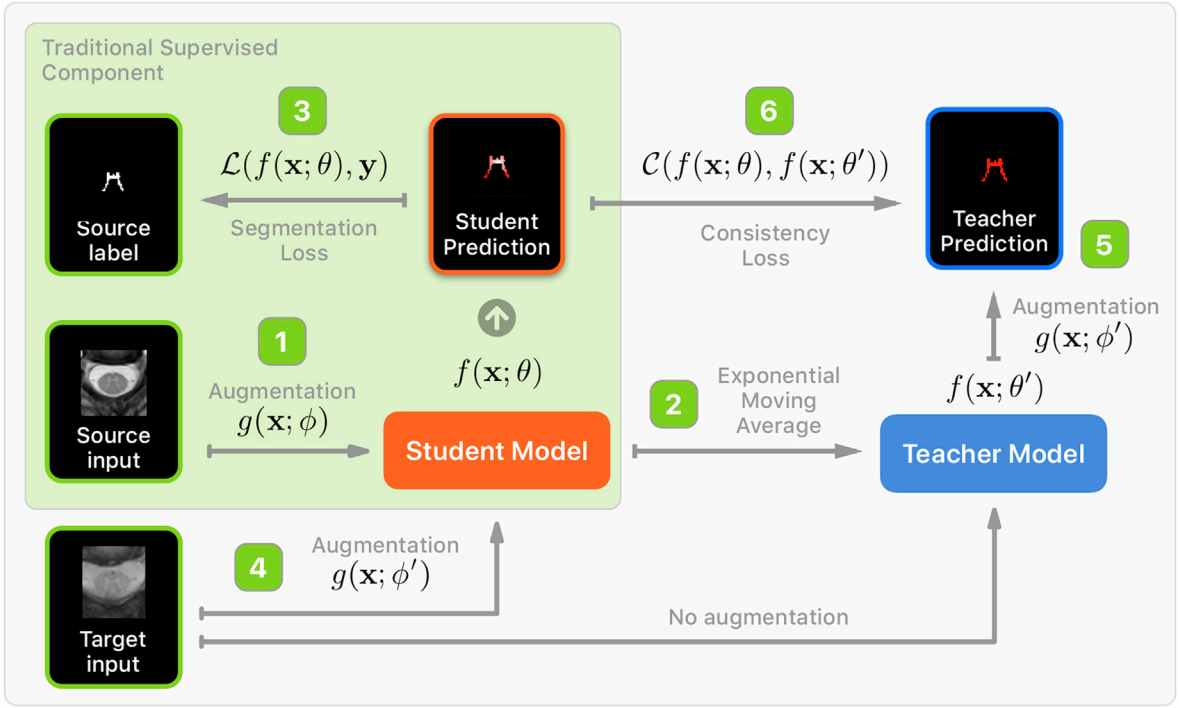

Fig. 4. Overview of the proposed method. The green panel represents the traditional supervision framework. **(1)** The source domain input slice data is augmented by the $g(x;\phi)$ transformation and fed into the student model. **(2)** The teacher model parameters is updated with an exponential moving average (EMA) from the student weights. **(3)** The traditional segmentation loss, where the supervision signal is provided with the source domain labels. **(4)** The input unlabeled slice data from the target domain is transformed with $g(x;\phi')$ before the student model forward pass (note the different parametrization ϕ'). **(5)** The teacher model prediction is transformed with $g(x;\phi')$ (same transformation as in Step 4). **(6)** The consistency loss, which enforces consistency between student and teacher predictions.

Table 1
Evaluation results in different centers. The evaluation and adaptation columns represent, respectively, the centers where testing and adaptation data were collected. Results are averages and standard deviations over 10 runs (with independent initialization of random weights). Values highlighted represent the best results at each center. All experiments were trained in both centers 1 and 2 simultaneously. The baseline rows are just the standard supervised learning procedure with centers 1 and 2 as training data and no additional information. Dice represents the Sørensen–Dice coefficient and mIoU represents the mean Intersection over Union — or Jaccard Index.

Evaluation	Adaptation	Dice	mIoU	Recall	Precision	Specificity	Hausdorff
Center 1	Baseline	47.25 ± 0.10	31.46 ± 0.08	**94.90**± 0.29	32.18 ± 0.09	99.66 ± 0.0	2.88 ± 0.01
	Center 3	47.71 ± 0.16	31.84 ± 0.14	94.18 ± 0.16	32.69 ± 0.15	99.67 ± 0.0	**2.85**± 0.01
	Center 4	**48.42**± 0.92	**32.47** ± 0.80	94.51 ± 0.57	**33.33**± 0.93	**99.68** ± 0.02	2.86 ± 0.02
Center 2	Baseline	50.69 ± 0.09	34.44 ± 0.08	**94.79**± 0.24	35.32 ± 0.10	99.61 ± 0.00	2.89 ± 0.01
	Center 3	51.05 ± 0.25	34.76 ± 0.23	93.78 ± 0.42	35.83 ± 0.31	99.62 ± 0.01	**2.87**± 0.01
	Center 4	**51.29**± 0.67	**34.98** ± 0.61	93.87 ± 0.91	**36.06**± 0.82	**99.63** ± 0.02	2.87 ± 0.02
Center 3	Baseline	82.81 ± 0.33	71.05 ± 0.36	**90.61**± 0.63	77.09 ± 0.34	99.86 ± 0.0	2.14 ± 0.02
	Center 3	**84.72**± 0.18	**73.67** ± 0.28	87.43 ± 1.90	**83.17**± 1.62	**99.91** ± 0.01	**2.01** ± 0.03
	Center 4	84.45 ± 0.14	73.30 ± 0.19	87.13 ± 1.77	82.92 ± 1.76	**99.91**± 0.01	2.02 ± 0.03
Center 4	Baseline	69.41 ± 0.27	53.89 ± 0.31	**97.22**± 0.11	54.95 ± 0.35	99.70 ± 0.00	2.50 ± 0.01
	Center 3	73.27 ± 1.29	58.50 ± 1.57	94.92 ± 1.48	60.93 ± 2.51	99.77 ± 0.03	2.36 ± 0.06
	Center 4	**74.67**± 1.03	**60.22** ± 1.24	93.33 ± 1.96	**63.62** ± 2.42	**99.80** ± 0.02	**2.29** ± 0.05

To answer *Question 2*, one can analyze the rows where both evaluation and adaptation centers are the same (3 or 4). Both rows present the highest values for almost all metrics (again, excepted for recall). This suggests that domain adaptation is working properly for that scenario.

Regarding *Question 3*, by looking at evaluation on center 3 and adaptation using center 4 (and vice-versa), we observe gains over the baseline once again for most metrics, suggesting that domain adaptation improves generalization for unseen centers.

6.2. Varying the consistency loss

We executed multiple runs of the Mean Teacher algorithm by varying the consistency loss to determine which one works best. We focused just on losses that do not contain additional hyperparameters. The Tversky Loss (Salehi et al., 2017), for instance, is quite similar to the Dice loss but with two additional hyperparameters (α and β).

Our choices of losses were thus limited to cross-entropy, mean squared error (MSE), and Dice, as previously described in Section 4. We believe, however, that a thorough analysis of distinct loss functions is of great importance for domain adaptation and should be explored in future work.

6.3. Behavior of dice loss and thresholding

A well-known fact regarding the Dice loss is that it usually produces predictions concentrated around the upper and lower bounds of the probability distribution, with very low entropy. As in (Perone and Cohen-Adad, 2018b), we used a high threshold value (0.99) for the Dice predictions to produce a balanced model. We have found, however, that the domain adaptation method also regularizes the network predictions, shifting the Dice probability distribution outside of the probability bounds. For that reason, we have decreased the Dice prediction threshold to 0.9 (instead of 0.99), which produced a more balanced model in terms of precision and recall.

6.4. Training stability

For unsupervised domain adaptation, it is important to have a stable training procedure. Since, in the most difficult scenarios, there are no annotations for validating the adaptation, an unstable training may produce sub-optimal adaptation results.

To evaluate the training stability, we tried distinct consistency weights for each possible consistency loss and we evaluated the difference between the best values that were found and the final results after 350 epochs. Table 2 summarizes results of this analysis. We also conducted experiments with an alternative formulation for dice loss where the denominator is the sum of the squared terms. We found that it heavily alleviated the problem of low-stability, but had poor results in terms of the dice score for every center, specially for Center 3.

We can observe that cross-entropy consistently fails, even with different weights, potentially due to the class imbalance of this particular task. Though it also achieves high dice values in its best scenario during training. Thus cross-entropy becomes a possible alternative to MSE when a few annotated images are available for validation in the target domain. Fig. 6 shows how the training diverges for cross-entropy after several iterations.

One way to alleviate this issue is to conduct an early stopping in the training. However, as we must assume that there are no labeled examples from the target center, the early stopping must be conducted with data from source centers. We investigated whether the epoch when drop in scores happen in the target center matches the one in source centers. We found that the drop somewhat appears at the same moment, but the

Table 2
Results on evaluating on center 3. The training set includes centers 1 and 2 simultaneously, with unsupervised adaptation for center 3. Values within parentheses represent the best validation results for each metric. The remaining values represent the final result after 350 epochs. CE is the cross-entropy loss, Dice represents the Sørensen–Dice coefficient, and MSE is the mean-squared error.

Loss	Weight	Dice	mIoU	Recall	Precision	Specificity	Hausdorff
CE	5	0.00 (85.50)	0.00 (74.91)	0.00 (95.01)	0.00 (98.90)	100.0 (100.00)	0.00 (0.00)
	10	0.00 (80.73)	0.00 (69.54)	0.00 (83.21)	0.00 (98.78)	100.0 (100.00)	0.00 (0.00)
	15	6.43 (37.03)	4.89 (26.06)	5.38 (77.05)	17.34 (65.85)	100.0 (100.00)	0.28 (0.00)
	20	2.30 (67.61)	1.86 (52.55)	2.09 (65.00)	7.94 (96.57)	100.0 (100.00)	0.12 (0.03)
Dice	5	76.76 (80.74)	62.76 (68.16)	97.88 (99.66)	63.72 (72.50)	99.71 (99.81)	2.36 (2.16)
	10	4.77 (10.55)	2.45 (5.64)	96.25 (99.99)	2.45 (5.85)	79.59 (99.75)	8.80 (2.57)
	15	2.30 (7.74)	1.16 (4.12)	99.95 (100.00)	1.16 (4.62)	55.07 (99.80)	11.75 (2.50)
	20	1.79 (4.43)	0.90 (2.27)	99.99 (100.00)	0.90 (2.30)	42.02 (99.84)	12.68 (2.43)
MSE	5	83.7 (83.88)	72.2 (72.46)	91.24 (98.19)	78.1 (78.57)	99.87 (99.93)	2.1 (2.00)
	10	84.38 (84.38)	73.19 (73.19)	90.15 (99.07)	80.12 (80.12)	99.88 (99.94)	2.05 (1.89)
	15	84.59 (84.59)	73.49 (73.50)	89.19 (98.52)	81.28 (81.28)	99.89 (99.89)	2.03 (2.03)
	20	84.5 (84.50)	73.36 (73.37)	90.36 (94.63)	80.16 (80.16)	99.88 (99.98)	2.05 (1.46)

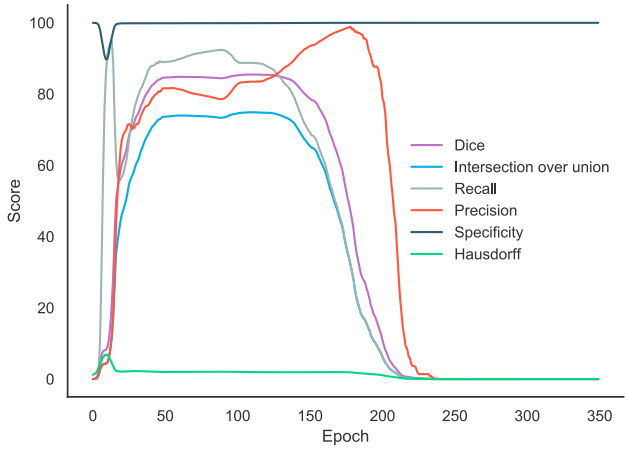

Fig. 6. Per-epoch validation results for the teacher model at center 3 with cross-entropy as the consistency loss. Training was conducted in both centers 1 and 2 simultaneously, and adapted to center 3 with consistency weight $\gamma = 5$.

target center dropped a bit earlier, see Fig. 7. This means that the early stopping should be conservative, choosing a model of several stages prior to when the drop in scores effectively occurred. We leave more of this discussion for future work.

We can observe that both Dice and cross entropy have trouble stabilizing the training after achieving high results. However, MSE tends to be more invariant to consistency weight, thus being a robust approach when no annotated data is available at the target center. As in (French et al., 2017), we also tried confidence thresholding, although we did not observe improvements.

7. Ablation studies

This section describes the ablation analyses, the purpose of which was to better understand the behavior of different components in the domain adaptation scenario.

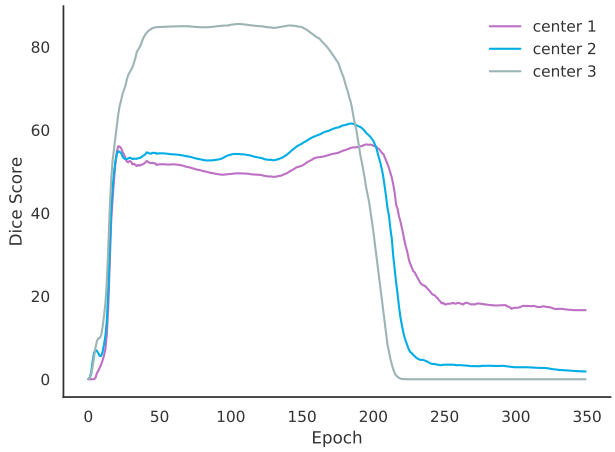

Fig. 7. Per-epoch dice score for the teacher model at center 3 with cross-entropy as the consistency loss. This is an evidence that early stopping can be conducted, although carefully, using data from the source centers. Training was conducted in both centers 1 and 2 simultaneously, and adapted to center 3 with consistency weight $\gamma = 5$.

7.1. Exponential moving average (EMA)

The improvement seen in Table 1 could also be explained by introducing the exponential moving average (EMA) during the training procedure, since it averages and smoothes the SGD trajectories.

To demonstrate that the improvement is specific to using unlabeled data and does not only come from the exponential average component, we performed an ablation experiment that leaves the EMA active but sets the consistency weight to zero. This experiment allowed us to evaluate the impact of the exponential average in the absence of the unlabeled data used to enforce consistency.

We reproduced the same experimental setup from Table 1 but with the consistency weight set to zero. Results are presented in Table 3 and show that the EMA model (teacher) is very similar to the baseline model. For every metric in the baseline we conducted a paired *t*-test, finding statistically significant results ($p < 0.03$) for dice, intersection over union, precision, and specificity metrics. Although statistically significant, there is only a small improvement, which could arguably be due to a poorly chosen α. However, note that Mean Teacher, which heavily relies on the EMA model, was nevertheless able to outperform a purely-supervised method by a great margin as seen in Table 1.

8. Domain shift visualization

Next, we investigated how domain adaptation affects the prediction space of segmentation at distinct centers. By using *t*-SNE (Maaten and Hinton, 2008), a non-linear dimensionality reduction technique, we were able to assess changes on the predictive perception of the network regarding unsupervised data. All data presented in the following figures were not used for training.

We created two baselines for this experiment. The first model was trained in a supervised fashion following the same hyperparameters presented in Section 4.4. The second was an adaptation scenario where both centers 1 and 2 were used as supervised centers and 3 as adaptation target. The vectors projected with *t*-SNE represents the features from the network prior to the final sigmoid activation.

Both *t*-SNE executions had a learning rate set to 10, perplexity to 30, and were executed for about 1000 iterations.[1] We notice that more iterations than 1000 preserved the groups' structure but further compressed them. This made visualizing the centers harder, so 1000 was a good trade-off between identifying emerging groups and interpretability.

Results from the supervised experiment are shown in Fig. 8a. Note that there is a clear separation between data from centers used during training (1 and 2) and unseen centers (3 and 4). This shows that the network predictions greatly differ according to the center to which the sample belongs to.

When adapting the network with unlabeled samples from a different domain, predictions become more diffuse, at least for centers presented during training. Results from the unsupervised adaptation experiment are shown in Fig. 8b. In that scenario, centers with labeled data (centers 1 and 2) form clusters with domains seen only in an unsupervised manner (3) or not presented to the network at all (4). A possible explanation for the change in clusters is that the model learns to map the manifold as it best suits for the task, instead of creating clusters of predictions based on whether the domain was *seen* or *unseen* during training.

9. Conclusion and limitations

Variability and scarcity of annotations in the medical imaging context is still challenging for machine learning. The large set of parameters that can be used to acquire image modalities and the lack of standardized protocols or industry standards are pervasive across the entire field.

[1] We used the TensorBoard embedding projector, available at https://github.com/tensorflow/tensorboard.

Table 3
Results of the ablation experiment where the baseline model was trained and compared against its exponential moving average (EMA) model without using Mean Teacher training scheme with unlabeled data. All experiments were trained in both center 1 and 2 simultaneously. Center 3 is the validation set and Center 4 is the test set. We also show the two-tailed p-value for each metric between the baseline and EMA models.

Evaluation	Version	Dice	mIoU	Recall	Precision	Specificity	Hausdorff
Center 3	Baseline	82.94 ± 0.35	71.20 ± 0.41	90.48 ± 0.45	77.39 ± 0.39	99.86 ± 0.00	2.13 ± 0.01
	EMA	82.97 ± 0.34	71.24 ± 0.40	90.51 ± 0.43	77.42 ± 0.40	99.86 ± 0.00	2.13 ± 0.01
	p-value	0.0024	0.0013	0.0429	0.0102	0.0201	0.5677
Center 4	Baseline	69.57 ± 0.22	54.08 ± 0.26	97.11 ± 0.12	55.20 ± 0.30	99.71 ± 0.00	2.50 ± 0.01
	EMA	69.59 ± 0.22	54.10 ± 0.26	97.09 ± 0.12	55.23 ± 0.31	99.71 ± 0.00	2.50 ± 0.01
	p-value	0.0743	0.0706	0.0773	0.0552	0.0627	0.3222

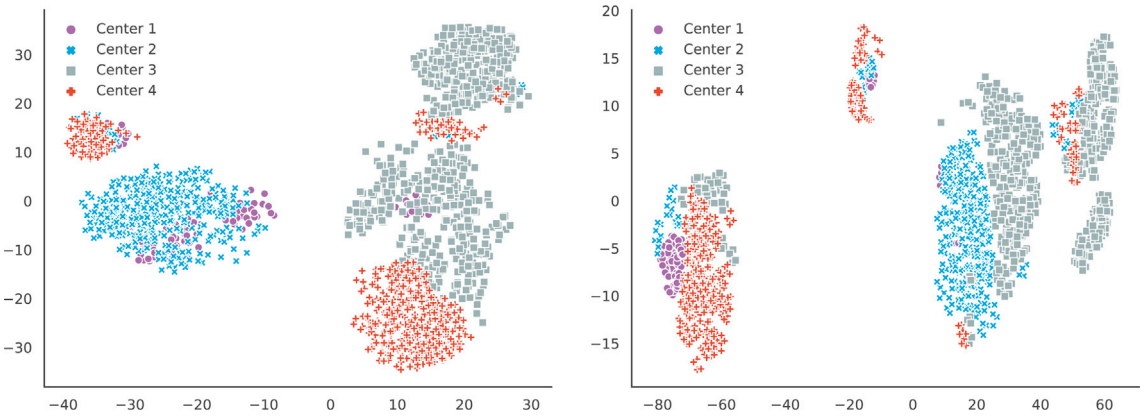

Fig. 8. Execution of t-SNE algorithm for two different scenarios. Colors represent data from different centers. a) A visualization of the t-SNE 2D non-linear embedding projection for the supervised learning scenario. The colors represent data from different centers. b) A visualization of the t-SNE 2D non-linear embedding projection for the domain adaptation scenario. The colors represent data from different centers.

In this work, we have shown that unsupervised domain adaptation, without depending on annotations, is an effective way to increase the performance of machine learning models for medical imaging across multiple centers.

Through the evaluation of multiple metrics in a large set of experiments, we have shown how self-ensembling methods can improve generalization on unseen domains through the leverage of unlabeled data from multiple domains. We also performed an ablation study that demonstrated strong evidence that the improvements come by the introduction of the unlabeled data and not only due to the exponential moving average.

We assessed how cross-entropy (when used as a consistency loss function) fails at maintaining training stability when the number of epochs progresses. We have discussed how this can lead to potential problems in more challenging scenarios for multiple centers. We also discussed issues related to the Dice loss when used as consistency loss.

We acknowledged the following limitations in our study. Firstly, we did not evaluate adversarial training methods for domain adaptation. Even considering the Mean Teacher as the current state-of-the-art method on many datasets, we believe that further analyses on the same realistic small data regime could significantly increase the importance of our contributions, and thus we leave that aspect for future work.

Secondly, the single-task evaluation of the gray matter segmentation could be extended to other tasks in other domains. Increasing the number of centers alongside the number of tasks would be relevant for confirming results obtained in the present study.

Further work on the field could lead to methods capable of measuring the risk of adaptation to particular centers or domains. This would be an important step towards understanding the limitations of the domain adaptation methods.

We believe that the problems that arise from the variability of medical imaging modalities require rethinking some of the strong assumptions made in machine learning models and training procedures. An important step in that direction is to reassess the importance of proper multi-domain evaluation in studies and medical imaging challenges, which rarely provide a test set from different domains (such as different centers, machines, etc) that contain the variability found in real-world scenarios.

10. Source-code and dataset availability

In the spirit of Open Science and reproducibility, the source-code used to perform the experiments presented in this study is publicly available.[2]

The dataset used for this work is also available on the Spinal Cord Gray Matter Segmentation Challenge website.[3]

Acknowledgments

We are very thankful to Ryan Topfer for the sensible review and time dedicated to improve this article. Funded by the Canada Research Chair in Quantitative Magnetic Resonance Imaging [950-230815], the Canadian Institute of Health Research [CIHR FDN-143263], the Canada Foundation for Innovation [32454, 34824], the Fonds de Recherche du Québec - Santé [28826], the Fonds de Recherche du Québec - Nature et Technologies [2015-PR-182754], the Natural Sciences and Engineering Research Council of Canada [435897-2013], the Canada First Research Excellence Fund (IVADO and TransMedTech) and the Quebec Bio-Imaging Network [5886]. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nivel Superior – Brasil (CAPES) – Finance Code 001.

[2] https://github.com/neuropoly/domainadaptation.
[3] http://cmictig.cs.ucl.ac.uk/niftyweb/program.php?p=CHALLENGE.

References

AlBadawy, E.A., Saha, A., Mazurowski, M.A., 2018. Deep learning for segmentation of brain tumors: impact of crossinstitutional training and testing. Med. Phys. 45 (3), 1150–1158.

Cao, J., Katzir, O., Jiang, P., Lischinski, D., Cohen-Or, D., Tu, C., Li, Y., 2018. Dida: Disentangled Synthesis for Domain Adaptation arXiv preprint arXiv:1805.08019.

Chen, C., Dou, Q., Chen, H., Heng, P.-A., 2018. Semantic-Aware Generative Adversarial Nets for Unsupervised Domain Adaptation in Chest X-Ray Segmentation arXiv preprint arXiv:1806.00600.

Coupé, P., Manjón, J.V., Fonov, V., Pruessner, J., Robles, M., Collins, D.L., 2011. Patch-based segmentation using expert priors: application to hippocampus and ventricle segmentation. Neuroimage 54 (2), 940–954.

Dou, Q., Ouyang, C., Chen, C., Chen, H., Heng, P.-A., 2018. Unsupervised Cross-Modality Domain Adaptation of Convnets for Biomedical Image Segmentations with Adversarial Loss (Tech. Rep.).

French, G., Mackiewicz, M., Fisher, M., 2017. Self-ensembling for Visual Domain Adaptation arXiv preprint arXiv:1706.05208.

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

Ghifary, M., Kleijn, W.B., Zhang, M., Balduzzi, D., Li, W., 2016. Deep reconstruction-classification networks for unsupervised domain adaptation. In: European Conference on Computer Vision, pp. 597–613.

Gros, C., De Leener, B., Badji, A., Maranzano, J., Eden, D., Dupont, S.M., et al., 2018, may. Automatic Segmentation of the Spinal Cord and Intramedullary Multiple Sclerosis Lesions with Convolutional Neural Networks arXiv preprint arXiv:1805.06349.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE International Conference on Computer Vision, 11–18-Dece, pp. 1026–1034. https://doi.org/10.1109/ICCV.2015.123.

Hinton, G., Vinyals, O., Dean, J., 2015. Distilling the Knowledge in a Neural Network arXiv preprint arXiv:1503.02531.

Hoffman, J., Tzeng, E., Park, T., Zhu, J.-Y., Isola, P., Saenko, K., et al., 2017. Cycada: Cycle-Consistent Adversarial Domain Adaptation arXiv preprint arXiv:1711.03213.

Hou, L., Samaras, D., Kurc, T.M., Gao, Y., Davis, J.E., Saltz, J.H., 2016. Patch-based convolutional neural network for whole slide tissue image classification. In: Proceedings of the Ieee Conference on Computer Vision and Pattern Recognition, pp. 2424–2433.

Ioffe, S., Szegedy, C., 2015. Batch normalization: accelerating deep network training by reducing internal covariate shift. In: Proceedings of the 32nd International Conference on International Conference on Machine Learning, vol. 37, pp. 448–456. https://doi.org/10.1007/s13398-014-0173-7.2.

Javanmardi, M., Tasdizen, T., 2018. Domain Adaptation for Biomedical Image Segmentation Using Adversarial Training. Isbi), pp. 554–558.

Kamnitsas, K., Baumgartner, C., Ledig, C., Newcombe, V., Simpson, J., Kane, A., et al., 2017. Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. In: International Conference on Information Processing in Medical Imaging, pp. 597–609.

Kingma, D.P., Ba, J.L., 2015. Adam: a method for stochastic optimization. In: International Conference on Learning Representations 2015, pp. 1–15. http://doi.acm.org.ezproxy.lib.ucf.edu/10.1145/1830483.1830503.

Lafarge, M.W., Pluim, J.P., Eppenhof, K.A., Moeskops, P., Veta, M., 2017. Domain-adversarial neural networks to address the appearance variability of histopathology images. In: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pp. 83–91. https://doi.org/10.1007/978-3-319-67558-9_10, 10553 LNCS.

Lai, M., 2015. Deep Learning for Medical Image Segmentation arXiv preprint arXiv:1505.02000.

Laine, S., Aila, T., 2016. Temporal Ensembling for Semisupervised Learning arXiv preprint arXiv:1610.02242.

LeCun, Y., Bengio, Y., Hinton, G., Y, L., Y, B., G, H., 2015. Deep learning. Nature 521 (7553), 436–444. https://doi.org/10.1038/nature14539.

Li, Y., Wang, N., Shi, J., Liu, J., Hou, X., 2016. Revisiting Batch Normalization for Practical Domain Adaptation arXiv preprint arXiv:1603.04779.

Lin, T., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal loss for dense object detection. IEEE International Conference on Computer Vision (ICCV), Venice 2999–3007. https://doi.org/10.1109/ICCV.2017.324.

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., et al., 2017. A survey on deep learning in medical image analysis. Med. Image Anal. 42, 60–88. https://doi.org/10.1016/j.media.2017.07.005.

Liu, Y.-C., Yeh, Y.-Y., Fu, T.-C., Wang, S.-D., Chiu, W.-C., Wang, Y.-C.F., 2018. Detach and adapt: learning cross-domain disentangled deep representation. In: Proceedings - 31th IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2018.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: Proceedings of the Ieee Conference on Computer Vision and Pattern Recognition, pp. 3431–3440.

Maaten, L. v. d., Hinton, G., 2008. Visualizing data using t-sne. J. Mach. Learn. Res. 9 (Nov), 2579–2605.

Madani, A., Moradi, M., Karargyris, A., Syeda-Mahmood, T., 2018. Semi-supervised learning with generative adversarial networks for chest x-ray classification with ability of data domain adaptation. In: IEEE 15th Symposium on Biomedical Imaging(Isbi), pp. 1038–1042. https://doi.org/10.1109/ISBI.2018.8363749.

Mahmood, F., Chen, R., Durr, N.J., 2018. Unsupervised reverse domain adaptation for synthetic medical images via adversarial training. IEEE Trans. Med. Imaging PP (c), 1. https://doi.org/10.1109/TMI.2018.2842767.

Milletari, F., Navab, N., Ahmadi, S.-A., 2016. V-net: fully convolutional neural networks for volumetric medical image segmentation. In: 3d Vision (3dv), 2016 Fourth International Conference on, pp. 565–571.

Neyshabur, B., Bhojanapalli, S., McAllester, D., Srebro, N., 2017. Exploring generalization in deep learning. In: Advances in Neural Information Processing Systems, pp. 5947–5956.

Odena, A., Oliver, A., Raffel, C., Cubuk, E.D., Goodfellow, I., 2018. Realistic Evaluation of Semi-supervised Learning Algorithms.

Oliver, A.G.B., Odena, A.G.B., Raffel, C.G.B., Cubuk, E.G.B., Goodfellow, I.J.G.B., 2018. Realistic evaluation of semi-supervised learning algortihms. In: International conference on Learning Representations, pp. 1–15.

Perone, C.S., Cohen-Adad, J., 2018a. Deep semi-supervised segmentation with weight-averaged consistency targets. DLMIA MICCAI, pp. 1–8. https://doi.org/10.1007/978-3-030-00889-5 sep.

Perone, C.S., Cohen-Adad, J., 2018b. Spinal cord gray matter segmentation using deep dilated convolutions. Nat. Sci. Rep. 8 (1) https://doi.org/10.1038/s41598-018-24304-3.

Polyak, B.T., Juditsky, A.B., 1992. Acceleration of stochastic approximation by averaging. SIAM J. Contr. Optim. 30 (4), 838–855.

Prados, F., Ashburner, J., Blaiotta, C., Brosch, T., Carballido-Gamio, J., Cardoso, M.J., et al., 2017. Spinal cord grey matter segmentation challenge. Neuroimage 152, 312–329. https://doi.org/10.1016/j.neuroimage.2017.03.010.

Rajpurkar, P., Hannun, A.Y., Haghpanahi, M., Bourn, C., Ng, A.Y., 2017. Cardiologist-Level Arrhythmia Detection with Convolutional Neural Networks (arXiv preprint).

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional Networks for Biomedical Image Segmentation, pp. 1–8. https://doi.org/10.1007/978-3-319-24574- 4_28 arXiv preprint arXiv:1505.04597.

Ruppert, D., 1988. E_cient Estimations from a Slowly Convergent Robbins-Monro Process (Tech. Rep.). Cornell University Operations Research and Industrial Engineering.

Salehi, S.S.M., Erdogmus, D., Gholipour, A., 2017. Tversky loss function for image segmentation using 3d fully convolutional deep networks. In: International Workshop on Machine Learning in Medical Imaging, pp. 379–387.

Salimans, T., Kingma, D.P., 2016. Weight normalization: a simple reparameterization to accelerate training of deep neural networks. In: Advances in Neural Information Processing Systems, pp. 901–909.

Sankaranarayanan, S., Balaji, Y., Castillo, C.D., Chellappa, R., 2018. Generate to adapt: aligning domains using generative adversarial networks. In: Proceedings - 31th IEEE Conference on Computer Vision and Pattern Recognition. https://doi.org/10.1109/CVPR.2017.316. CVPR 2018.

Santurkar, S., Tsipras, D., Ilyas, A., Madry, A., 2018. How Does Batch Normalization Help Optimization? (No, it Is Not about Internal Covariate Shift) arXiv preprint arXiv:1805.11604.

Sun, B., Saenko, K., 2016. Deep coral: correlation alignment for deep domain adaptation. In: European Conference on Computer Vision, pp. 443–450.

Tarvainen, A., Valpola, H., 2017. Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results. In: Advances in Neural Information Processing Systems, pp. 1195–1204.

Tzeng, E., Ho_man, J., Zhang, N., Saenko, K., Darrell, T., 2014. Deep Domain Confusion: Maximizing for Domain Invariance arXiv preprint arXiv:1412.3474.

Wang, M., Deng, W., 2018. Deep Visual Domain Adaptation: A Survey. Neurocomputing.

Wu, Y., He, K., 2018. Group Normalization arXiv preprint arXiv:1803.08494.

Yosinski, J., Clune, J., Bengio, Y., Lipson, H., 2014. How transferable are features in deep neural networks? Adv. Neural Inf. Process. Syst. 27, 1–9 (Proceedings of NIPS), 27.

Zamir, A.R., Sax, A., Shen, W., Guibas, L.J., Malik, J., Savarese, S., 2018. Taskonomy: disentangling task transfer learning. In: The Ieee Conference on Computer Vision and Pattern Recognition (Cvpr).

Zech, J.R., Badgeley, M.A., Liu, M., Costa, A.B., Titano, J.J., Oermann, E.K., 2018. Confounding Variables Can Degrade Generalization Performance of Radiological Deep Learning Models arXiv preprint arXiv:1807.00431.

Zhang, Y., Miao, S., Mansi, T., Liao, R., 2018. Task Driven Generative Modeling for Unsupervised Domain Adaptation: Application to X-Ray Image Segmentation, vol. 2, pp. 1–9. https://doi.org/10.1007/978-3-030-00934-2_67 arXiv preprint arXiv:1806.07201.

Zhong, E., Fan, W., Yang, Q., Verscheure, O., Ren, J., 2010. Cross validation framework to choose amongst models and datasets for transfer learning. In: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 6323. LNAI, pp. 547–562. https://doi.org/10.1007/978-3-642-15939-8.

Zhu, J.-Y., Park, T., Isola, P., Efros, A.A., 2017. Unpaired Imagetoimage Translation Using Cycle-Consistent Adversarial Networks (arXiv preprint).