Domain Adaptation with Missing Data for Medical Image Segmentation

by

Mathilde BATESON

MANUSCRIPT-BASED THESIS PRESENTED TO ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
IN PARTIAL FULFILLMENT FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
Ph.D.

MONTREAL, JUNE 23 2023

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

BOARD OF EXAMINERS

THIS THESIS HAS BEEN EVALUATED

BY THE FOLLOWING BOARD OF EXAMINERS

Mr. Ismail Ben Ayed, Thesis Supervisor
Department of Electrical Engineering, ÉTS Montréal, Canada

Mr. Hervé Lombaert, Thesis Co-supervisor
Department of Software and IT Engineering, ÉTS Montréal, Canada

Mr. Carlos Vasquez, Chair, Board of Examiners
Department of Electrical Engineering, ÉTS Montréal, Canada

Mr. Marco Pedersoli, Member of the Jury
Department of Electrical Engineering, ÉTS Montréal, Canada

Mr. Ender Konukoglu, External examiner
ETH Zurich, Switzerland

THIS THESIS WAS PRESENTED AND DEFENDED

IN THE PRESENCE OF A BOARD OF EXAMINERS AND THE PUBLIC

ON APRIL 28 2023

AT ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

Adaptation au domaine avec données manquantes pour la segmentation d'images médicales

Mathilde BATESON

RÉSUMÉ

Les méthodes d'adaptation au domaine (DA) ont récemment attiré l'attention dans le contexte de la vision par ordinateur, car elles améliorent la transférabilité des modèles de réseaux profonds d'un domaine source à un domaine cible présentant des caractéristiques différentes. Le DA est essentiel pour atténuer le besoin d'annotations laborieuses requises par les modèles de segmentation profonde. Cependant, le cadre usuel des méthodes de DA n'est pas réaliste, car il nécessite l'accès à plusieurs ensembles de données, à la fois dans le domaine source et dans le domaine cible. Or, en milieu clinique, seuls quelques échantillons cibles, voire un seul, sont généralement disponibles, tandis que les données sources peuvent être inaccessibles. Par conséquent, l'objectif principal de cette thèse est de proposer des algorithmes de DA pour segmenter des images médicales avec des jeux de données d'entraînement limités.

Dans notre premier objectif, nous explorons le DA des réseaux de segmentation avec des annotations minimales dans le domaine cible. Nous abordons le DA par la segmentation sous des contraintes d'inégalité sur les prédictions des échantillons cibles non annotés ou faiblement annotés. Ainsi, nous faisons correspondre implicitement les statistiques de prédiction des domaines cible et source, avec une incertitude autorisée des connaissances préalables. Nous abordons le problème d'optimisation sous contrainte qui en découle avec des pénalités différentiables, parfaitement adaptées aux approches conventionnelles de descente de gradient stochastique.

Dans notre deuxième objectif, nous introduisons le DA sans source pour la segmentation d'images. Notre formulation est basée sur la minimisation d'une fonction d'entropie définie sur les données du domaine cible, que nous guidons en outre avec un a priori sur les régions de segmentation. Un a priori de ratio de classe est estimé à partir des connaissances anatomiques et intégré sous la forme d'une divergence de Kullback-Leibler dans notre fonction de coût globale. Nous montrons l'efficacité de notre méthode dans une variété de scénarios de DA, avec différentes modalités et applications, notamment la segmentation de la colonne vertébrale, de la prostate et du cœur.

Dans notre troisième objectif, nous étudions une méthode de DA sans source à utiliser au moment du test avec un seul sujet cible. Nous étudions un objectif de minimisation de l'entropie guidé par des a-priori de formes pour adapter la segmentation au moment du test. Nous explorons le potentiel de l'intégration de diverses contraintes sous la forme de moments de forme, pour guider le DA vers des solutions plausibles. En particulier, nous exploitons la taille, le centroïde et la distance au centroïde des structures anatomiques par le biais de contraintes de pénalité dans notre fonction de perte globale. Notre méthode est validée dans deux tâches de segmentation difficiles : L'adaptation IRM-CT pour des images cardiaques et l'adaptation inter-sites pour des images de la prostate.

En conclusion, chaque objectif pousse progressivement plus loin la complexité de la tâche d'adaptation et la quantité de données manquantes, afin d'obtenir un cadre clinique réaliste. Les méthodes proposées relèvent ce défi en étudiant comment exploiter au mieux les a priori du domaine, tels que la connaissance des formes anatomiques.

Mots-clés: Adaptation au domaine, segmentation sémantique, optimisation sous contrainte, apprentissage profond, imagerie médicale

Domain Adaptation with Missing Data for Medical Image Segmentation

Mathilde BATESON

ABSTRACT

Domain Adaption (DA) methods have recently attracted substantial attention in computer vision as they improve the transferability of deep network models from a source to a target domain with different characteristics. DA is key in mitigating the need for laborious pixel annotations required by deep segmentation models.

However, the framework of most common DA methods is not realistic, as it requires access to whole datasets, both in the source and in the target domain. Yet in clinical settings, only a few or even a single target sample(s) are typically available, while the source data might be inaccessible.

Therefore, this thesis main objective is to propose domain adaptation algorithms to segment medical images with limited training datasets.

In our *first* objective, we explore adapting segmentation networks with inequality constraints on the network predictions of target samples. Thereby, we implicitly match the prediction statistics of the target and source domains, with permitted uncertainty of prior knowledge. We address the ensuing constrained optimization problem with differentiable penalties, fully suited for conventional stochastic gradient descent approaches.

In our *second* objective, we introduce source-free domain adaptation for image segmentation. Our formulation is based on minimizing a label-free entropy loss defined over target-domain data, which we further guide with a domain-invariant class-ratio prior on the segmentation regions. The prior is estimated from anatomical knowledge and integrated in the form of a Kullback–Leibler divergence in our overall loss function. We show the effectiveness of our prior-aware entropy minimization in various source-free domain adaptation scenarios, with different modalities and applications, including spine, prostate and cardiac segmentation.

In our *third* objective, we study a source-free adaptation method for use at test-time with a single target subject. We investigate shape-guided entropy minimization objectives. We explore the potential of integrating various constraints in the form of shape moments, to guide domain adaptation towards plausible solutions. In particular, we exploit the size, centroid, and distance-to-centroid of anatomical structures through penalty constraints in our overall loss function. In our applications, an estimation of these shape moments is derived from textbook anatomical knowledge. Our method is validated in two challenging source-free single-subject adaptation tasks: MRI-to-CT adaptation for cardiac images, and cross-site adaptation for prostate images. The efficiency of 2D and 3D shape constraints are demonstrated in both applications.

In conclusion, each objective progressively pushes further the complexity of the adaptation task and the amount of missing data, to achieve a realistic clinical setting. The proposed methods

address this challenge by studying how to best leverage domain knowledge such as anatomical shape knowledge.

This thesis led to six different publications as first author, including three at MICCAI conferences, two journal publications, one in IEEE Transactions for Medical Imaging and one in Medical Image Analysis (MedIA), and one ongoing submission to MedIA. All the codes ensuing from this thesis are publicly available, and free to reuse and modify. The functional programming style used makes it easy to integrate new loss functions and shape information, with little-to-no additional coding efforts.

Keywords: Domain Adaptation, Semantic segmentation, Constrained optimization, Deep learning, Medical imaging

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CT	Computed tomography
MRI	Magnetic resonance imaging
ML	Machine Learning
DL	Deep Learning
CV	Computer Vision
DNN	Deep neural network
CNN	Convolutional neural network
GD	Gradient Descent
SGD	Stochastic Gradient Descent
DA	Domain Adaptation
TTA	Test-Time Adaptation

LIST OF SYMBOLS

The following list of symbols is consistent thorough most of this thesis, although it might differ slightly in some papers. In such occasions, the notations re-introduced in the paper will take precedence.

$\Omega \subset \mathbb{R}^{2,3}$	Image space
$\mathcal{K} = \{1, \ldots, K\}$	Discrete set of labels
M	Number of modalities (channels) in an image
$\mathcal{N}(\cdot; \boldsymbol{\theta})$	Neural network with parameters $\boldsymbol{\theta}$
$\mathcal{D} = \{(x_n, y_n)\}_{n=1}^{N}$	Dataset
$x_n : \Omega \to \mathbb{R}^M$	Input image of sample n
$y_n : \Omega \to \{0, 1\}^K$	One-hot encoded segmentation mask label for sample n
$p_n(\boldsymbol{\theta}) : \Omega \to [0, 1]^K$	Softmax output for $\mathcal{N}(x_n; \boldsymbol{\theta})$

INTRODUCTION

Artificial intelligence (AI) bears many promises for medicine. First, computer-assisted medicine holds the potential to transform clinical practice, improving health outcomes for patients with less burden on the healthcare system. Second, AI could become a cornerstone of medical research, leading to breakthroughs in the understanding of complex diseases, and accelerating the development of new drugs.

Computer vision (CV) systems aim to make sense of the visual world automatically, mimicking and surpassing the human visual system. The breakthroughs in deep learning (DL) research and development Krizhevsky, Sutskever & Hinton (2012); Le, Monga, Devin, Corrado, Chen, Ranzato et al. (2013) led to significant progress in the CV domain. Thanks to advances in highly parallelizable Graphical Processing Units (GPUs) and the open-sourcing of large annotated datasets of natural images (such as ImageNet Russakovsky, Deng, Su, Krause, Satheesh, Ma et al. (2015)), DL-based methods have become the de-facto choice in handling many CV tasks. In this context, there is a widespread hope that similar progress will unfold in the medical imaging field, where adoption has been slower due to characteristics of medical imaging datasets and tasks that we will describe in the next section.

0.1 Computer vision for medical imaging: challenges and opportunities

0.1.1 Contemporary medical imaging

Advancements in cutting-edge research for image acquisition techniques and decreasing costs of storage solutions have substantially increased the size and number of digital images in radiological imaging departments. The most common contemporary medical imaging methods include:

- *Magnetic Resonance Imaging (MRI).* MRI is a medical imaging technique that can generate high contrast images of various soft tissues and organs, and is routinely used in many diagnostic studies of the head, spine, and joints. At the base of MRI is a property of

atoms: nuclear magnetic resonance (NMR), a physical phenomenon in which nuclei in an external magnetic field absorb and re-emit electromagnetic radiation. More specifically, the source of MRI is the hydrogen proton spin and its associated dipole magnetic moment. The composition of the scanned area is obtained by measuring the change in response, and the time it takes for the molecules to relax. Several MRI modalities exist, which can produce images of different aspects. Therefore, acquiring images across different MRI modalities is often very useful to capture different physical properties. Amongst the drawbacks of MRI imaging are the long acquisition time and the sensitivity to patient motion, especially at higher resolutions. Considerable research efforts are continuously improving speed, resolution, and patient comfort.

- *Computer tomography (CT).* The basis for X-ray imaging is the body's ability to absorb X-rays as they travel through the patient. Each body tissue has a specific absorption and radiation profile, letting through a specific amount of X-rays. Computed tomography (CT) imaging is a cross-sectional imaging technique: a rotating X-Ray machine performs a series of 2D scans at different angles, followed by the 3D reconstruction of the image. The main drawback of CT is the radiation dose that the patient receives, limiting the frequencies at which it can be used. Amongst its advantages, a CT scan is performed in minutes, making it an excellent tool in emergency situations.

- *Positron emission tomography (PET).* Positron emission tomography is a nuclear medicine imaging procedure that involves injecting small amounts of radioactive material which will accumulate in tumors or regions of inflammation, allowing the detection of diseases earlier than other imaging tests. PET is often used in conjunction with other diagnostic tests, such as CT, to produce more precise information and a more exact diagnosis. The main disadvantage of PET is the small amount of radiation that remains in the body. Additionally, a PET scan can take up to several hours, depending on the patient's condition.

- *Ultrasound Imaging (UI).* Ultrasound imaging is based on the emission of sound waves, which are reflected at the acoustic boundary between tissues (e.g., between different organs). Measuring the reflected waves allows to reconstruct the tissue layouts. Although it is the

most accessible of the medical imaging techniques, and allows real-time imaging, its main drawback is the low resolution and noise disturbance of the images obtained. However, its safety and cost-effectiveness makes it an attractive choice for first screening. Common procedures include abdominal ultrasound, bone sonometry, breast ultrasound, Doppler (to visualize blood flow), fetal ultrasound, and echocardiogram.

The choice of imaging modality used in the clinical practice depends on the body part that needs to be analysed, the safety concerns regarding radiation exposure, and the cost and availability of the machines.

Along with these specialized medical imaging techniques, devices such as smartphones and wearables track, analyze and store massive amounts of data, creating a personalized “data-ome". There is great hope that integrating all such data will fuel a transformation of healthcare, from treatment to prevention. However, along with these opportunities comes the increasing need for secure data management, and recommendations for the safe adoption of DL methods in the medical field Acosta, Falcone, Rajpurkar & Topol (2022). We will review such challenges in the next section.

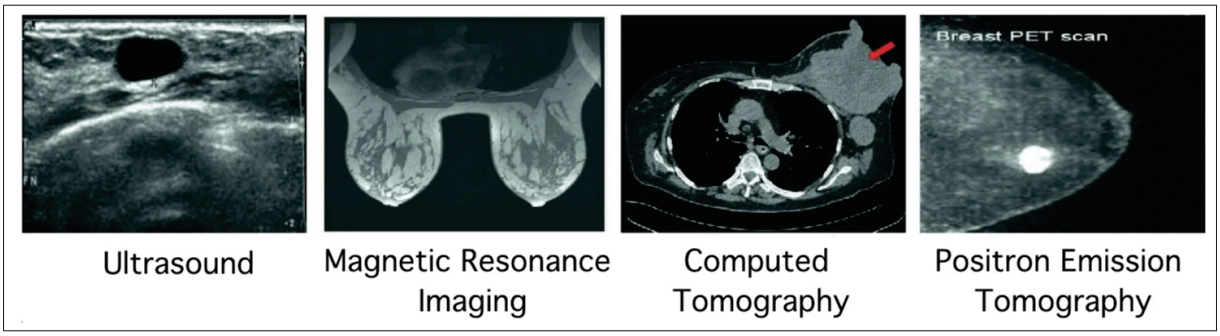


Figure 0.1 Example of medical image modalities

0.1.2 Challenges, limitations, and methodological recommendations for the safe adoption of deep-learning enabled CV

0.1.2.1 Characteristics of medical imaging datasets

Medical images have a few characteristics which set them apart from datasets in other vision fields. As a result, medical image datasets available for learning to perform a task are often subjected to specific difficulties summarized below.

0.1.2.1.1 Data availability and quality

Creating medical imaging datasets involves numerous technical tasks including acquisition, reconstruction, enhancement, and restoration. Artifacts arise at each step, such as respiratory motion, bias fields, shadowing, signal dropout. A second challenge is that recent developments in medical imaging acquisition techniques have substantially increased the size of digital images. A raw medical image can take more than 1 GB ! This limits the possibility of sharing raw medical images between healthcare information systems. Digital compression can help for archiving and transmission of medical images. But the lossy compression on which most efficient techniques rely is avoided in practice, due to the risk of losing critical information which could create legal ramifications. Additionally, most medical images are in 3D, forcing standard models to either work with a set of 2D slices, or adjust their internal structure to process in 3D. Moreover, although radiological imaging departments have seen an increasing amount of digital images, few of these can actually be used for machine learning tasks. First, strict regulations regarding the sharing of medical images often impede using them, even by practitioners in the clinical entity where they were acquired. Second, even when sharing is allowed, medical images are expensive and hard to interpret. This work is traditionally done by trained radiologists. To accurately label medical images, various medical experts are commonly required to minimize human bias. Otherwise, labels are considered noisy. Despite the latest advances in CV, annotating medical images is still largely carried out manually, often by creating a pixel-level mask manually. Hence, a critical characteristic of medical images is that their **annotations are extremely expensive to**

make. As a result, medical imaging datasets are usually much smaller than in other computer vision fields: whereas the ImageNet dataset has more than 1 million images with bounding box annotations, most initiatives to gather annotated data in medical images do not exceed a few thousand images.

0.1.2.1.2 It's not all about larger datasets

Even when available, larger datasets have not always led to the progress hoped for. For instance, Varoquaux & Cheplygina (2022) showed that an increase in dataset size was not met with increased accuracy to predict pathological versus stable evolution for patients with Alzheimer's disease. In fact, larger studies tended to report worse prediction accuracy, which is worrisome as they are closer to real-life settings. We identify two major underlying causes:

- *Biases in datasets available for learning and testing:* To produce clinically-relevant predictions, the learning and testing data must match the actual target population, which is often not the case when datasets are small. As a result, algorithms that score high in benchmarks can perform poorly in real world scenarios. Such bias has been demonstrated in retinal imaging Tasdizen, Sajjadi, Javanmardi & Ramesh (2018), chest X-rays Zech, Badgeley, Liu, Costa, Titano & Oermann (2018), brain MRIs Wachinger, Rieckmann & Pölsterl (2021), and dermatology Abbasi-Sureshjani, Raumanns, Michels, Schouten & Cheplygina (2020). Addressing this bias would require a better standardization of data gathering protocols. However, a great part of medical data will always be captured opportunistically, using equipment, patients and specialists when available. Quantifying and alleviating these biases is therefore a difficult task in many applications.

- *Data imbalance:* One of the inherent problems related to CV and ML in general is data imbalance. This refers to a big discrepancy in distribution between classes of a dataset, with one class being several orders of magnitude more frequent than another. Moreover, oftentimes, the class of interest (presence of disease, malignant tumor...) is much less frequent than the "contrast" class (healthy, benign). If training methods do not explicitly

account for this imbalance, the resulting predictions will overestimate the dominant class and underestimate the minority classes, which can have severe consequences : this is called a learning bias.

On the other hand, medical images present some advantages compared to natural images. First, they tend to be more constrained, with stable texture and shape of anatomical structures - brains do not twist or make faces, as humans or cats do! In many cases, some typical problems such as viewpoint are less critical, as medical images are often acquired in standard orientations (axial, coronal, and sagittal planes).

0.1.2.2 Methodological recommendations

Despite promising in silico results, many of the CV approaches are not adopted in the clinical practice. While the reasons behind this can be complicated and varied, the failure to adapt and to be robust in clinical practice is often cited as the main cause. Comparing to the widely used Technology Readiness Levels (TRL) framework for systems engineering Lavin *et al.* (2022) (see Figure 0.2), the authors of Vlontzos *et al.* (2022) showed that medical imaging CV applications frequently skip from TRL 4 (proof of concept) to TRL 7 (deployment), ignoring the crucial TRLs 5 and 6, which make new systems resilient to real-world conditions.

For the safe adoption of computer vision systems in the clinical practice, and the acceleration of the TRL1-TRL6 pipeline, the following methodological recommendations related to the robustness of CV models have been outlined Esteva, Chou, Yeung, Naik, Madani, Mottaghi et al. (2021); Litjens et al. (2017):

- *Regularization strategy:* Without abundant training samples, as is common in medical applications, a DL model faces the challenge of overfitting, whereby its high capacity memorizes particularities of the training set which prevent it from generalizing to new data. Overfitting can be monitored by measuring the gap between the training error and validation error, where overfitting corresponds to a large gap. Regularization encompasses

Figure 0.2 Technology Readiness Levels of ML Systems. Medical CV applications often skip levels TRL5 and TRL6 where algorithms are made robust and production ready. Image sources: Lavin *et al.* (2022) (top); Vlontzos *et al.* (2022) (bottom)

any modification made to a learning algorithm in order to reduce its validation error but not its training error.

- *Interpretability of results:* Explainable AI aims to render model behavior understandable by humans, allowing domain experts to detect and prevent shortcut learning arising from spurious correlations. For example, accidentally fitting confounders has been shown in various disease detection tasks, with AI models more likely to predict the presence of the disease when the image contained a ruler Esteva, Kuprel, Novoa, Ko, Swetter, Blau et al. (2017), skin markings Winkler, Fink, Toberer, Enk, Deinlein, Hofmann-Wellenhof et al. (2019), an 'urgent' mark Badgeley, Zech, Oakden-Rayner, Glicksberg, Liu, Gale et al. (2019), or the use of a portable X-ray machine Zech *et al.* (2018). Understanding the features learned

by neural networks is a first step towards the ultimate goal of extracting causal relations from correlative patterns. Explainable AI is especially valuable in high-risk settings, and will facilitate the wide acceptance of automated decisions in the medical community. Explaining the outcomes of CV algorithms for medical imaging is challenging due to the complexity of data and models. A few recent works on interpretable CV models for medical imaging have emerged Kowarsch, Weijler, Wödlinger, Reiter, Maurer-Granofszky, Schumich et al. (2022); Sun, Darbehani, Zaidi & Wang (2020a); Zhou, Guo, Shen & Yang (2020). For instance, to explain semantic segmentation predictions, saliency methods are often adopted, but several important limitations have been identified in Saporta, Gui, Agrawal, Pareek, Truong, Nguyen et al. (2022). In short, deep learning explainability for medical imaging tasks is still nascent.

- *Generating fair and unbiased results:* Developing more accurate algorithms is the main focus of efforts from the medical image computing research community. However, the global accuracy of models can hide blind spots resulting from societal biases present in the data. Algorithmic unfairness can be caused by (1) data selection bias, such as under-representation of minorities; (2) model selection bias towards accuracy on the majority at the detriment of some subgroups, and (3) outcome noise due to unobserved variables Kelly, Karthikesalingam, Suleyman, Corrado & King (2019). Beyond global accuracy, efforts should be made towards producing AI systems that perform well regardless of gender, ethnicity, race, patient diagnosis, etc Obermeyer & Topol (2021). Similarly to explainable AI, the field of fairness in medical imaging AI is still in its infancy Lara, Echeveste & Ferrante (2022).

- *Robustness to domain shifts:* The diversity in acquisition settings (vendors, sites), directly affects the performances of AI systems. Even small changes between settings might cause a model trained on one setting to fail on a different one. Because of the aforementioned constraints on medical imaging, obtaining a training dataset covering several sites, vendors and settings is often impossible. Learning models robust to distribution shifts is the goal of Domain Adaptation. **This will be the main focus of this dissertation.**

0.1.3 Potential impacts of Computer Vision in medical imaging

There is growing evidence that deep neural networks can be trained to "interpret" medical images Esteva *et al.* (2021). This holds implications in many clinical tasks, such as screening, diagnosis, predicting future outcomes, monitoring disease, and clinical research. The following CV tasks are central to realizing this vision and have led to an extensive amount of research on machine learning for medical images :

- *Image registration* or correspondence establishes which parts of one image correspond to which parts of another image or model. Active challenges include the registration of brains with malformations in their development Shuvaev, Lazutkin, Kiryanov, Anokhin, Enikolopov & Koulakov (2022).

- *Image classification* consists in assigning a label to a whole image, such as a diagnosis in medical imaging. An example of a difficult problem is the classification of a tumor as benign or malignant Yap, Yolland & Tschandl (2018).

- *Semantic segmentation* is the most challenging task as it aims at pixel-level prediction, and requires expensive pixel-level annotations for training. It is of great value for the diagnosis, treatment and follow-up of many diseases. Amongst hard segmentation tasks is the segmentation of moving organ parts, such as the left ventricle in Cine-MRI Oksuz, Mukhopadhyay, Dharmakumar & Tsaftaris (2017), which is helpful in diagnosing cardiac arrhythmia. Automating segmentation would be greatly beneficial in many fields such as oncology, where it is routinely performed, as it allows to discriminate between the areas to radiate and those to spare during radiotherapy Mirikharaji & Hamarneh (2018). Illustrations of medical image segmentation are shown in Figure 0.3. **This is the application we emphasize in this PhD dissertation.**

After the automatic interpretation of medical images, computer assistance can be extended to downstream tasks. For instance, computer vision models are being developed to predict the survival to cancer Chen, Lu, Weng, Chen, Williamson, Manz et al. (2021b), to help in the

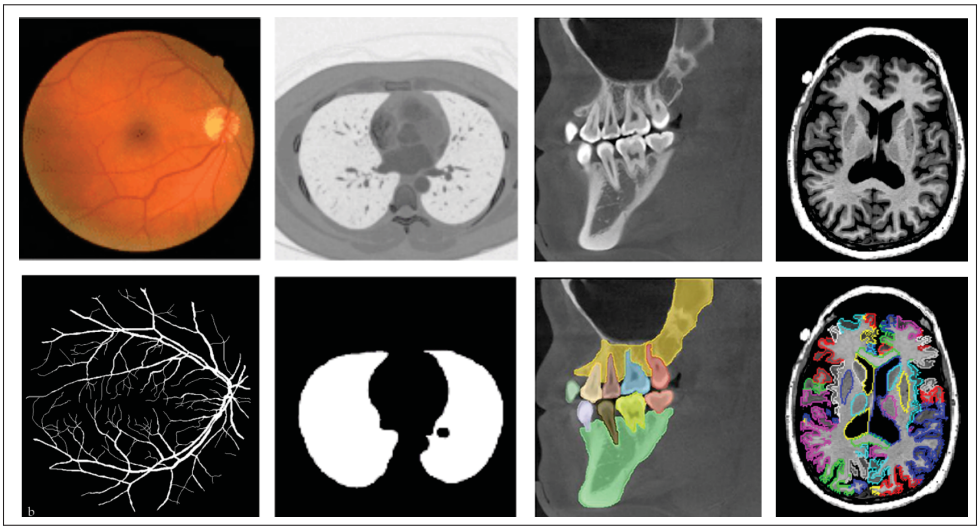


Figure 0.3 Illustrations of medical images (top) and their associated semantic segmentation (bottom)

treatment decision, or to help uncover disease characteristics that cannot be assessed by the naked eye. A revealing example is that of the retina, with increasing evidence indicating that it contains rich information that humans, including experts, cannot see.

0.1.3.1 Motivating example : what AI models see in the retina

A first surprising result is that while ophthalmologists cannot determine gender from retinal images, as there are no cues visible to human eyes, a recent deep learning model obtained 97% accuracy for gender determination Poplin, Varadarajan, Blumer, Liu, McConnell, Corrado et al. (2018). As expected, AI models are quite efficient for detecting eye diseases, such as diabetic retinopathy. However, less obvious results include the detection of kidney disease Sabanayagam, Xu, Ting, Nusinovici, Banu, Hamzah et al. (2020) and hepatobiliary disease Xiao, Huang, Wang, Lin, Zhu, Chen et al. (2021); the control of blood sugar and blood pressure Poplin *et al.* (2018); the prediction of heart attack Diaz-Pinto, Ravikumar, Attar, Suinesiaputra, Zhao, Levelt et al. (2022); the close correlation of the retinal vessels with the cardiac arterial calcium score Rim, Lee, Tham, Cheung, Yu, Lee et al. (2021); the ongoing prospective assessment and tracking of Alzheimer's disease Wagner, Hughes, Cortina-Borja, Pontikos, Struyven, Liu et al. (2022).

Illustrations of these promising results are seen in Figure 0.4. These reports have been published in the last few years, and more discoveries will surely follow regarding what CV models can "see" in retinal images that would be invisible to human eyes. Therein lies the future potential to obtain an accurate readout of many organ functions via a simple smartphone photo of the retina.

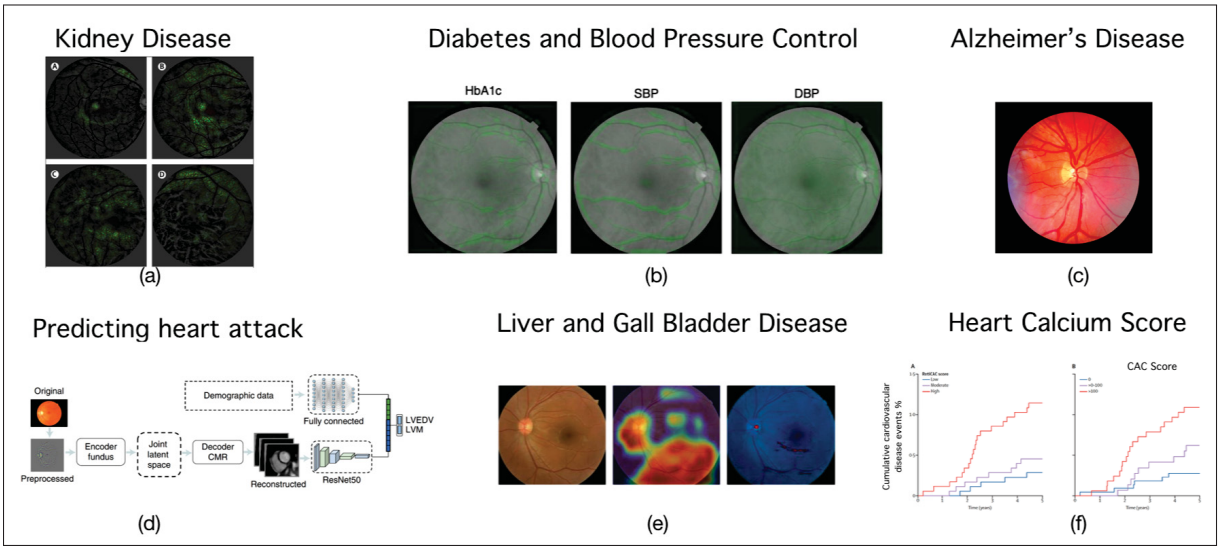


Figure 0.4 Examples of the capability of computer vision models to "see" well beyond human limits in retinal imaging. Figures from Sabanayagam *et al.* (2020) (a); Poplin *et al.* (2018) (b) ; Wagner *et al.* (2022) (c); Diaz-Pinto *et al.* (2022) (d); Xiao *et al.* (2021) (e); Rim *et al.* (2021) (f)

Without denying such advances, a key limitation is that these studies are based on retrospective analyses. In silico analysis of complete, "cleaned" datasets are useful for hypotheses generation. To ensure clinical trust, they should be confirmed by rigorous prospective and/or randomized clinical trials that validate AI algorithms in real-world clinical settings Kelly *et al.* (2019).

0.2 Motivations and objectives

Many of the challenges encountered when applying computer vision models to medical imaging are related to data availability and quality. While scaling up datasets can help, ensuring strictly controlled and consistent conditions for data gathering is not always possible: in the medical

domain, some data will always need to be collected opportunistically Elyan, Vuttipittayamongkol, Johnston, Martin, McPherson, Jayne et al. (2022). Hence, overcoming the challenges due to limited data also calls for the design of adequate learning methods to make the most use of CV models in medical imaging. In this context, this PhD dissertation addresses Domain Adaptation (DA), i.e. the transferability of a model trained on a source domain to new target domains. The motivation of this work starts from the observation that the setting of most common DA methods is not realistic, as it requires access to whole datasets, both in the source and in the target domain. However, in clinical settings, only a few or even a single target sample is typically available, while the joint use of source and target dataset might be denied, for privacy reasons Kaissis, Makowski, Rückert & Braren (2020).

The research questions we would like to address are the following : Can we rely on target domain knowledge such as shape information to guide the adaptation of deep segmentation models, and can we integrate these priors directly into the optimization loss El Jurdi, Petitjean, Honeine, Cheplygina & Abdallah (2021) ? Can we forego access to the source data in the adaptation phase of a deep segmentation model, to comply with a more realistic clinical setting ? Can we adapt segmentation networks when only a few (or even a single) samples are provided in the target domain ?

Therefore, the main objective of this thesis is to propose domain adaptation methods guided by target domain knowledge to adapt deep segmentation models images with limited training datasets. There is strong motivation to use prior knowledge directly into the adaptation framework: why re-learn (through expensive, annotated data) information readily available in textbooks, or in radiology reports ?

In our *first* objective, we introduce a general novel methodology based on constrained domain adaptation for image segmentation, incorporating prior knowledge about the segmentation regions. We impose inequality constraints on the network predictions of unlabeled or weakly labeled target samples. We address the ensuing constrained optimization problem with differentiable penalties, fully suited for conventional stochastic gradient descent approaches. The comparison

with state-of-the-art adaptation methods reveals a considerably better performance of our model on two challenging tasks. Our results also show robustness to imprecision in the prior knowledge.

In our *second* objective, we introduce source-free domain adaptation for image segmentation. Our formulation is based on minimizing a label-free entropy loss defined over target-domain data, which we further instruct with a class-ratio prior on the segmentation regions. The class-ratio prior is estimated from anatomical knowledge and incorporated via a Kullback–Leibler (KL) divergence in our global loss function. Moreover, we motivate our optimization loss by linking it to the maximization of the mutual information between the target images and their label predictions. We show the effectiveness of our prior-aware entropy minimization in various domain-adaptation scenarios, with distinct modalities and applications, including spine, prostate and cardiac segmentation.

In our *third* objective, we investigate a source-free adaptation method for use at test-time with a single target subject. We investigate a shape-guided entropy minimization objective for fine-tuning segmentation networks at test-time. We explore the potential of integrating various constraints in the form of shape moments, to guide domain adaptation towards plausible solutions. In particular, we exploit the size, centroid, and distance-to-centroid of anatomical structures through penalty constraints in our overall loss function. In our applications, an estimation of these shape moments is derived from textbook anatomical knowledge. Our method is validated in two challenging segmentation tasks: MRI-to-CT adaptation for cardiac images, and cross-site adaptation for prostate images. The efficiency of 2D and 3D shape constraints is demonstrated in both applications. Our approach allows the integration of different annotation levels in a new target domain, from no supervision to weak supervision that informs on shape moments. It consistently yields significantly better results than existing test-time adaptation methods. More surprisingly, it outperformed some state-of-the-art domain adaptation methods, even though it foregoes both training on additional target data during adaptation and access to the source data.

In conclusion, each objective progressively pushes further the complexity of the adaptation task and the amount of missing data, to achieve a realistic clinical setting. The proposed methods

address this challenge by studying how to best leverage domain knowledge such as anatomical shape knowledge, as well as general-purpose concepts from the self-supervised field, such as Shannon entropy minimization. In the following chapters, we will show that this allows to adapt segmentation models with less data, while leading them towards more robust, and realistic predictions.

0.3 Thesis outline

Chapter 1 presents a short review of neural networks for semantic segmentation, where we cover the standard segmentation training losses and models. We then present domain adaptation for image segmentation, the shortcomings of their traditional setting, and a few alternatives proposed in the literature to achieve adaptation.

Chapter 2 presents our first research objective, a domain adaptation method based on a constrained formulation, embedding domain-invariant prior knowledge about the segmentation regions. Our general formulation imposes inequality constraints on the network predictions of the target samples, and address the ensuing constrained optimization problem with differentiable penalties, fully suited for conventional stochastic gradient descent approaches. This chapter corresponds to the paper entitled "Constrained domain adaptation for image segmentation", published in IEEE Transactions for Medical Image Analysis.

Chapter 3 presents our second research objective, a source-free domain adaptation for image segmentation. To guide adaptation in this context, our formulation is based on minimizing a label-free entropy loss defined over target-domain data. Contrasting with Chapter 1, we combined entropy minimization with a domain-invariant prior, in the form of a class-ratio prior on the segmentation regions. The work in this chapter is published in the Journal of Medical Image Analysis, titled "Source-Free domain adaptation for image segmentation".

Chapter 4 presents our third research objective, a single-subject test-time adaptation framework from segmentation. We explore the potential of combining entropy minimization with various constraints in the form of shape moments, to guide domain adaptation towards plausible solutions.

In particular, we exploit the size, centroid, and distance-to-centroid of anatomical structures through penalty constraints in our overall loss function. The content of this chapter corresponds to the paper "Single-Subject Test-Time adaptation with shape moments for image segmentation" submitted to the Journal of Medical Image Analysis.

The **Conclusion** chapter summarizes each contribution for the three research objectives with its practical impact, current limitations and possible directions for future works.

0.4 Published Work

The findings in this thesis have led to the following publications.

- **M. Bateson**, J. Dolz, H. Kervadec, H.Lombaert, I. Ben Ayed. "Constrained domain adaptation for image segmentation" *MICCAI 2019, journal extension in IEEE TMI, volume 40, 2021*.
- **M. Bateson**, H. Kervadec, J. Dolz, H.Lombaert, I. Ben Ayed. "Source-Free domain adaptation for image segmentation", *MICCAI 2021, journal extension in MedIA, volume 83, 2022.*
- **M. Bateson**, H.Lombaert, I. Ben Ayed. "Test-Time adaptation with shape moments for image segmentation", *MICCAI 2022, journal extension ongoing revision in MedIA*.

0.5 Code and open-source

The code of all papers is available, free to reuse/modify. While split in different repositories, the code stems from the same (private) codebase, that expanded over the years of this PhD.

Constrained domain adaptation for image segmentation

https://github.com/mathilde-bat/CDA

Source-Free domain adaptation for image segmentation

https://github.com/mathilde-bat/SFDA

Test-Time adaptation with shape moments for image segmentation

https://github.com/mathilde-bat/TTA

CHAPTER 1

BACKGROUND

1.1 Optimization and learning: High-level overview

We start by reviewing *gradient-based optimization*, a fundamental tool for deep learning.

1.1.1 Gradient-based optimization

Continuous optimization consists in finding the optimal value (either a maximum or minimum) of a function $F : \mathbb{R}^D \to \mathbb{R}$ with respect to its input $x \mapsto F(x)$. It has widespread real-world applications, as many problems can be formulated as a continuous optimization problem. Given the following minimization problem:

$$\min_{x \in \mathbb{R}^D} F(x). \tag{1.1}$$

A global optimal solution $p^* := F(x^*)$ will verify:

$$\forall y \in \mathbb{R}^D : p^* \leq f(y), \tag{1.2}$$

while a locally optimal $\hat{p} := f(\hat{x})$ for a neighborhood $\Omega \in \mathbb{R}^D$ will verify:

$$\forall y \in \Omega : \hat{p} \leq f(y), \tag{1.3}$$

Finding the *global* optimal p^* (or the corresponding optimal input x^*) can be very difficult, and often cannot be solved analytically. However, when $F \in C_1(\mathbb{R}^d)$, finding a local optimum can be done by the gradient descent (GD) algorithm, a central optimization approach described in Algorithm 1.1.

The motivation for GD is rooted in the observation that the gradient $\nabla F(x)$ is the *slope* of the function at that point. The GD method consists in starting with an initial guess x^0, and then following this slop until finding a local optimum. When F is convex, we have the guarantee that a local optimum is a *global* optimum: $\hat{x} = x^*$. In the case of a non-convex function, the simple gradient-based optimization can get *stuck* in a local optimum and more complex schemes are needed to find the global optimum.

Algorithm 1.1 Overview of the gradient descent (GD) algorithm

Input: *Given a step size* τ*, Given stopping criterion* ϵ
Output: *Current solution* $\hat{x} := x^t$
Initialize x^0 to some value in $\mathbb{R}^D$
Initialize $t \leftarrow 0$
while ϵ *is not met* **do**
    $x^{t+1} \leftarrow x^t - \tau \nabla F(x^t)$
    $t \leftarrow t + 1$
end

Optimization methods for convex problems are quite mature and robust, with known convergence properties Boyd & Vandenberghe (2004). On the contrary, optimizing non-convex problem such as those encountered in machine learning algorithms is much more complex, and the guarantees from convex optimization are mostly lost. Learning is still possible though, as we discuss in the next section.

1.1.2 Machine learning differs from pure optimization

A machine learning (ML) algorithm is an algorithm that is able to learn from data to perform a task. This is observed when given a performance measure P, its performance at the task as measured by P improves with processing the data.

When training a ML model, we have access to a set of N examples from the data-generating distribution $\mathcal{D} = \{x_n\}_{n=1}^N$. To increase the performance P, an objective function $\mathcal{L}$ with parameters $\theta \in \mathbb{R}^D$ is computed and minimized on this set. A differentiable function of is

typically chosen for $\mathcal{L}$, and gradient-based optimization such as Algorithm 1.1 is favoured to minimize $\mathcal{L}$. What we have so far described is merely an optimization problem. What distinguishes ML from pure optimization, where minimizing $\mathcal{L}$ is a goal in and of itself, is that we want the generalization error, i.e. the error on a new unseen test set to be low as well. Regularization is a key idea in machine learning that refers to a variety of methods for modifying an algorithm to reduce its generalization error.

The objective function of ML algorithms $\mathcal{L}$ usually decomposes as a sum over the training examples:

$$\min_{\theta} \sum_{x_n \in \mathcal{D}} \mathcal{L}(x_n; \theta) \tag{1.4}$$

Algorithm 1.1 is usually avoided for ML algorithms. Indeed, it would require evaluating the gradient $\nabla\mathcal{L}(x_n; \theta)$ on every example in the entire training dataset, which can become very expensive. A modification of the GD is used in nearly all of ML instead: the stochastic gradient descent (SGD), which relies on taking the average gradient on a mini-batch B of examples drawn i.i.d from the distribution $\mathcal{D}$. The algorithm is briefly described in Algorithm 1.2.

Algorithm 1.2 Overview of the stochastic gradient descent algorithm for ML

Input: *Given a step size τ, a mini-batch size B, a stopping criterion ϵ (convergence, or quality of the result), a distribution Π, an uniform distribution U,*
Input: *Given a set of N examples from the distribution $\mathcal{D} = \{x_n\}_{n=1}^{N}$*
Input: *Given a differentiable objective function $\mathcal{L}$*
Output: *Current solution $\hat{\theta} := \theta^t$*
Initialize $t \leftarrow 0$
while *ϵ is not met* **do**
    Sample $\mathcal{B} \sim U(0, N)^B$
    $L = \sum_{b \in \mathcal{B}} \mathcal{L}((x_b; \theta))$
    $\theta^{t+1} := \theta^t - \tau\nabla L$
    $t \leftarrow t + 1$
end

Note that it is also common to choose an adaptive step size τ_t instead of the fixed one in Algorithm 1.2, and is the basis of popular algorithms derived from SGD, such as Adam Kingma & Ba (2014).

In addition to these practical considerations, employing SGD rather than GD is recommended since the noise introduced to the learning process can have a regularizing effect. Moreover, the SGD estimation of the gradient is unbiased.

Regularization can also help express preferences for specific solutions of Eq. (1.4), both implicitly and explicitly. For instance, a ML model can be regularized by adding a penalty called a regularizer R to the objective function:

$$J(\theta) = \sum_{(x_n)\in\mathcal{D}} \mathcal{L}(x_n;\theta) + \lambda R(\theta) \tag{1.5}$$

Minimizing J results in a choice of parameters θ that make a trade-off between minimizing the objective function L and the regularizer.

For instance, the regularizer R is often designed to express a generic preference for simpler models in order to promote generalization. A common such regularizer in ML and DL is *weight decay*, which expresses a preference for smaller weights θ explicitly, via $R(\theta) = \theta^T\theta$.[1]

Alternatively, some regularization strategies put extra constraints and penalties on a ML model, designed to encode specific kinds of prior knowledge. We will see examples of such regularization techniques for segmentation models in Section 1.3.4.

[1] In ML models, there are often two types of parameters: the weights and the biases. The weights directly influence the relationship between the inputs and the outputs of the model because they are multiplied by the inputs. The biases only offset the relationship from the intercept. Therefore only the weights are usually regularized.

1.2 Deep neural networks

1.2.1 High-level overview

Deep learning belongs to representation learning methods, where a model automatically discovers useful representations from raw data for the desired task. DL methods are based on multiple levels of abstraction, built by composing simple but non-linear processing *layers*. Each layer transforms the representation at one level (starting with the raw data) into a representation at a higher and more abstract level. By composing sufficient layers, very complicated features may be learned.

Formally, a neural network $\mathcal{N}$ can be written as a function composition of its layers l_i:

$$\mathcal{N}(x;\boldsymbol{\theta}) := l_p \circ ... \circ l_0(x), \tag{1.6}$$

Each l_i is an instance of standard operations, represented by its parameters:

$$\begin{aligned} w_i, b_i &:= \boldsymbol{\theta}_i \\ l_i\left(x;(w_i,b_i)\right) &:= g(xw_i + b_i), \end{aligned} \tag{1.7}$$

where g is a non-linear, derivable function. Some layers are much simpler, and are integrated to reduce the dimensionality of the representation (by max-pooling for instance). Let $\mathcal{D} = \{(x_n, y_n)\}_{n=1}^{N}$ be a set of inputs and their corresponding labels. For classification tasks, y_n takes a value among a set of discrete labels $\mathcal{K} = \{1, ..., K\}$. From the network output $\mathcal{N}(\cdot;\boldsymbol{\theta}) \in \mathbb{R}^K$, the class $\hat{y}_n$ predicted by the network is obtained with :

$$\hat{y}_n = \arg\max_{k\in\mathcal{K}} \mathcal{N}(x_n;\boldsymbol{\theta})_k. \tag{1.8}$$

A particular type of deep network, deep convolutional neural networks (CNN), were introduced to process multi-array data, such as images in computer vision. In order to find a specific pattern

in a whole array, CNN use a convolution operation, inspired by signal processing. This consists in carrying out the same operation on subsets of the input, in a sliding window fashion. In CNN, this introduces the prior that useful features are translation equivariant, and allows for a great reduction in the number of model parameters. CNN led to spectacular improvement in vision recognition problems LeCun, Bengio & Hinton (2015).

Once the deep network architecture is defined (the function composition $\mathcal{N}$), its parameters or weights[2] $\boldsymbol{\theta}$ need to be tuned for the network to perform well on the desired task. This was seen for a long time as an intractable problem, as $\mathcal{N}$ is a very non-convex and high dimensional function, and was the main cause of the unpopularity of deep networks. The big breakthrough came with the discovery that deep network architectures could be trained by the simple SGD described in Algorithm 1.2 Bottou & Bousquet (2007). Although no theoretical results regarding convergence or optimality are guaranteed, it was confirmed experimentally that CNN could be very trained via the gradient-based backpropagation procedure described below, as long as the modules were smooth enough functions of their inputs and of their weights Hardt, Recht & Singer (2016).

1.2.2 Learning the weights of a deep network

In practice, given an underlying data generating distribution $\mathfrak{p}(x, y) \in \mathfrak{p}$, a labeled dataset $\mathcal{D}$ is then sampled i.i.d. from $\mathfrak{p}(x, y) : \mathcal{D} = \{(x_n, y_n)\}_{n=1}^{N}$. "Learning" a deep neural network model is a term used to describe the following process : a batch of B samples are passed to the network, which outputs its predictions on this batch; a loss is computed and used to modify the network parameters by iterating through a forward and a backward process described below. A number of training iterations over the whole training dataset, or epochs, is necessary to obtain a good performance.

In order to optimize the network parameters $\boldsymbol{\theta}$, a loss function $\mathcal{L}$ is introduced. It is usually designed to be minimal only when the network predictions match perfectly the ground truth labels y_n, and to be an increasing function of the mismatch. This loss is optimized with respect

[2] The term weights is a misnomer as technically neural networks parameters are divided between weights and the biases, but is widely used as a synonym in the ML community.

to the network parameters $\boldsymbol{\theta}$:

$$\arg\min_{\boldsymbol{\theta}} \sum_{(x_n, y_n) \in \mathcal{D}} \mathcal{L}(\mathcal{N}(x_n; \boldsymbol{\theta}), y_n). \tag{1.9}$$

We summarize below a few of the building blocks of neural network optimization.

1.2.2.1 Softmax

The arg max function from Eq. (1.8) is not derivable, and is therefore incompatible with gradient descent. From the raw network outputs $z = \mathcal{N}(x_n; \boldsymbol{\theta})$, continuous probabilities are obtained with the derivable softmax function:

$$p_n^k(\boldsymbol{\theta}) := \frac{e^{z_k}}{\sum_{j=1}^{K} e^{z_j}} \text{ , k=1...K} \tag{1.10}$$

The resulting $\left(p_n^k(\theta)\right)_{k=1...K}$, often called the softmax output or simply the softmax, is a vector of the continuous probability predictions of the network ($\sum_{k=1}^{K} p_n^k(\boldsymbol{\theta}) = 1$). An exact solution would predict a probability of 1 for the class y_n, and 0 for all others.

1.2.2.2 Backpropagation

Backpropagation is the standard procedure to train deep networks. It is a practical application of the chain rule for derivatives, used to compute:

$$\frac{\partial \mathcal{L}(\mathcal{N}(x_n, \theta), y_n)}{\partial \boldsymbol{\theta}} \tag{1.11}$$

It indicates how a multilayer model should change its internal parameters $\boldsymbol{\theta}_l$ from layer l from the representation in the subsequent layer $l + 1$. The backpropagation algorithm is applied repeatedly to propagate gradients through all modules, starting from the output at the top (where the network produces its prediction) all the way to the bottom (where the external input x is fed).

1.2.2.3 Batch Normalization

Batch Normalization (BN) layers Ioffe & Szegedy (2015) have become standard components in deep networks. BN normalizes inputs to have zero-mean and unit variance and then transforms these "whitened" activations using affine parameters γ and β to maintain expressiveness. BN is well-known to smooth the optimization landscape, to speed up the training of CNN with stochastic gradient descent, and to enhance model performance Ioffe & Szegedy (2015); Li, Wang, Shi, Liu & Hou (2016); Nado, Padhy, Sculley, D'Amour, Lakshminarayanan & Snoek (2020).

Formally, given a batch size B and $\mathbf{x} \in \mathbb{R}^{H \times W \times B}$ activations in each channel, BN is expressed as

$$\mathrm{BN}(\mathbf{x}[i,j,b];\gamma,\beta) = \gamma \cdot \hat{\mathbf{x}}[i,j,b] + \beta, \tag{1.12}$$

where

$$\hat{\mathbf{x}}[i,j,b] = \frac{\mathbf{x}[i,j,b] - \mu}{\sqrt{\sigma^2 + \epsilon}}. \tag{1.13}$$

The mean and variance of activations within a mini-batch, μ and σ, are given by

$$\begin{aligned} \mu &= \frac{\sum_b \sum_{i,j} \mathbf{x}[i,j,b]}{B \cdot H \cdot W} \\ \sigma^2 &= \frac{\sum_b \sum_{i,j} (\mathbf{x}[i,j,b] - \mu)^2}{B \cdot H \cdot W}. \end{aligned} \tag{1.14}$$

During training, BN estimates the mean and variance of the activations, denoted by $\bar{\mu}$ and $\bar{\sigma}$, by exponential moving average. Given the t^{th} mini-batch, the mean and variance are estimated as:

$$\begin{aligned} \bar{\mu}^{t+1} &= (1-\alpha)\bar{\mu}^t + \alpha\mu^t, \\ \left(\bar{\sigma}^{t+1}\right)^2 &= (1-\alpha)\left(\bar{\sigma}^t\right)^2 + \alpha\left(\sigma^t\right)^2. \end{aligned} \tag{1.15}$$

In the testing phase, the mean and variance estimations $\bar{\mu}$ and $\bar{\sigma}$ are used to normalize input activations. However, it should be noted that if the testing images do not come from the

distribution $\mathfrak{p}$, sharing the mean and variance between the training and testing phase can be inappropriate.

1.3 Neural networks for semantic segmentation

Semantic segmentation refers to the partitioning of an image into multiple segments/regions. To this end, a label is assigned to every pixel such that pixels with the same label share similar properties. Semantic segmentation can also be seen as pixel-level classification.

Let us first define $\Omega \subset \mathbb{R}^{2,3}$ the image spacial domain of our dataset $\mathcal{D} = \{(x_n, y_n)\}_{n=1}^{N}$. x_n is an input image and y_n its corresponding one-hot encoded ground truth. For semantic segmentation, the network architecture is designed in such a way that its output matches the dimension of the inputs:

$$\begin{aligned} \mathbf{x}_n &: \Omega \rightarrow \mathbb{R}^M \\ \mathbf{y}_n &: \Omega \rightarrow \{0,1\}^K \\ \mathbf{p}_n(\boldsymbol{\theta}) &: \Omega \rightarrow [0,1]^K, \end{aligned} \tag{1.16}$$

where $\mathbf{p}_n(\boldsymbol{\theta}) = [\mathbf{p}_n(i, \boldsymbol{\theta}) \; \forall i \in \Omega]$ is the softmax output tensor, and M represents the number of channels (3 in the case of RGB images, 1 for grayscale images, the most common case in medical imaging) of the input.

Once the network weights $\boldsymbol{\theta}$ have been optimized, the final predicted segmentation is obtained by:

$$\begin{aligned} \hat{\mathbf{y}}_n &:= [\hat{\mathbf{y}}_n(i) \; \forall i \in \Omega] \\ \hat{\mathbf{y}}_n(i) &:= \arg\max_{k \in K} \mathbf{p}_n^k(i, \boldsymbol{\theta}) \end{aligned} \tag{1.17}$$

1.3.1 Performance metrics of segmentation models

We list below the most common evaluation metrics used to measure the quality of a segmentation model, by comparing its predictions with the ground-truth labels. Evaluation metrics can be divided into two categories : counting metrics or overlap-based metrics (DSC and IoU), for measuring the overlap between the ground-truth labels and the predictions; and distance-based metrics, for measuring the distance between object boundaries (HD, ASSD):

DSC index (F1 score) :

$$DSC(A, B) = \frac{2|\mathrm{A} \cap \mathrm{B}|}{|\mathrm{A}| + |\mathrm{B}|} \tag{1.18}$$

IoU (Jaccard Index) :

$$IoU(A, B) = \frac{|\mathrm{A} \cap \mathrm{B}|}{|\mathrm{A} \cup \mathrm{B}|} \tag{1.19}$$

Hausdorff Distance (HD) :

$$HD(A, B) = \max \left\{ \max_{a \in A} d(a, B), \max_{b \in B} d(A, b) \right\} \tag{1.20}$$

where $d(a, B) = \min_{b \in B} d(a, b)$

A common variant of the HD, the HD95, calculates the 95% percentile instead of the maximum, thereby disregarding outliers. Another popular metric is the Average Symmetric Surface Distance (ASSD), measuring the average of all distances for every point from one object to the other and vice versa:

ASSD :

$$ASSD(A, B) = \frac{\sum_{a \in A} d(a, B) + \sum_{b \in B} d(A, b)}{|A| + |B|} \tag{1.21}$$

Each individual performance metric suffers from limitations, such as their behavior in the presence of class imbalance or small target structures, which may bias the comparison of

segmentation models. Therefore, it is good practice to use two or more complementary metrics. For a thorough review of these limitations, we refer to Reinke, Eisenmann, Tizabi, Sudre, Rädsch, Antonelli et al. (2021). Note that the metrics can be computed either in 2D and or 3D, depending on the application.

After presenting the common performance metrics of segmentation models, we will describe in the next section common *supervised* losses for semantic segmentation, i.e. in the setting where the labels $(\mathbf{y}_n)_{n\leq N}$ are available.

1.3.2 Training supervised segmentation models

The following section presents the standard supervised-learning losses used in image segmentation. As they are averaged over the current mini-batch b, for readibility reasons we will write $\mathbf{y}_n$ as y, $\mathbf{p}_n(\boldsymbol{\theta})$ as $p_{\boldsymbol{\theta}}$, $\mathbf{y}_n(i)$ as $y(i)$, $\mathbf{p}_n(i,\boldsymbol{\theta})$ as $p(i,\boldsymbol{\theta})$, $y_n^k(i)$ as $y^k(i)$, and $p_n^k(i,\boldsymbol{\theta})$ as $p^k(i)$.

1.3.2.1 Common supervised segmentation losses

Ideally, we would like to directly improve one or more of the prediction performance metrics described in Section 1.3.1. However, as the arg max Eq. (1.8) is non-differentiable, this makes them impossible to incorporate directly in the gradient-based training of CNN. Hence, surrogate differentiable objective functions are designed, with the hope that minimizing them will improve the chosen performance metric(s).

Amongst useful properties of an objective function, we highlight (i) **a stable gradient** to ensure a smooth optimization via SGD; (ii) **calibration to the evaluation metric** – to ensure that a surrogate is useful, it needs to relate to the performance metric in some ways. The calibration to the evaluation metric captures this property, ensuring that a solution to the surrogate problem is also a solution to the original performance score maximization problem. (iii) **confidence calibration** a good match between the confidence of a model and its correctness - as miscalibration makes network predictions hard to rely on (iv) **robustness to class imbalance**, i.e. when very different segmentation targets region sizes are involved. In segmentation problems,

susceptibility to class imbalance can result in a collapse of predictions to the dominant class, often the background class. Ideally, a loss should have all of the characteristics. Unfortunately, as we will see below, in the most common supervised segmentation losses, there is a trade-off between these properties.

Cross-entropy loss: Cross-entropy or negative log likelihood is a standard quality measure of a probabilistic model. CE measures the difference between the label distribution and the predicted distribution. It takes the following form :

$$\mathcal{L}_{\mathrm{CE}}(y, p_{\boldsymbol{\theta}}) = -\sum_{i \in \Omega} \sum_{k} y^k(i) \log p^k(i) \tag{1.22}$$

The minimum $\mathcal{L}_{\mathrm{CE}}(y, p_{\boldsymbol{\theta}}) = 0$, corresponds to $p_{\boldsymbol{\theta}} = y$.

Note that CE loss is computed as the average of per-pixel error, without knowledge of the adjacent pixels' prediction. Because CE does not consider prediction errors globally, it is vulnerable to class-imbalance, leading to an over-segmentation of larger classes, and under-segmentation of smaller ones. Indeed, the CE loss is calibrated to the accuracy (which is not an appropriate metric in class imbalanced situations) and not to the more common DSC coefficient. Another well-known issue of the CE loss is the overconfidence in its prediction, leading to miscalibration Liu, Ben Ayed, Galdran & Dolz (2022), as CE encourages the predicted softmax probabilities $p_{\boldsymbol{\theta}}$ to match the one-hot label assignments y. Still, CE is favoured in many applications, because of its nice gradient profile allowing good training stability.

Weighted Cross-entropy loss: WCE is a variant of CE which introduces class weights to increase the contributions of the minority classes, thereby alleviating class imbalance.

$$\mathcal{L}_{\mathrm{WCE}}(y, p_{\boldsymbol{\theta}}) = -\sum_{i \in \Omega} \sum_{k} \nu_k y^k(i) \log p^k(i) \tag{1.23}$$

where $\nu_k, k = 1, \ldots, K$, are non-negative constants denoting class weights.

Focal loss: The Focal loss Lin, Goyal, Girshick, He & Dollár (2018) is a variant of the Cross-Entropy loss that addresses the issue of class imbalance by down-weighting the contribution of easy examples (for which the network predicts a high probability for the correct class), thereby focusing learning from the harder examples.

$$\mathcal{L}_{\mathrm{F}}(y, p_{\boldsymbol{\theta}}) = -\sum_{i \in \Omega} \sum_{k} y^k(i)(1 - p^k(i))^{\gamma} \log p^k(i) \tag{1.24}$$

Specifically, the focal loss causes the gradient norms for confident samples to be lower than they would have been with cross-entropy. Note that the focal loss is dependent on a hyperparameter γ.

Dice loss: As the DSC is the common performance metric taken to measure the quality of a segmentation model, it motivated Milletari, Navab & Ahmadi (2016) to introduce a soft-DSC surrogate, i.e. a differentiable loss that would be calibrated to this measure. The DSC formulation is relaxed to use the predicted continuous probabilities $p_{\boldsymbol{\theta}}$ instead of binary labels :

$$\mathcal{L}_{DSC}(y, p_{\boldsymbol{\theta}}) = \sum_{k=1}^{K} -\frac{2 \sum_{i \in \Omega} p^k(i) y^k(i)}{\sum_{i \in \Omega} p^k(i) + \sum_{i \in \Omega} y^k(i)} \tag{1.25}$$

Different from the previous per-pixel losses which are distribution-based, the Dice loss is geometry-based, and is computed from both local errors (numerator) and the global error (denominator). This attention to geometry allows the Dice loss to yield improvements for imbalanced segmentation tasks compared to the CE loss, as observed in a variety of medical applications Milletari *et al.* (2016). For a theoretical perspective on this observation, we refer to Liu, Dolz, Galdran, Kobbi & Ayed (2021a), where a hidden label-marginal bias of Dice towards extremely imbalanced solutions is uncovered. Note that a caveat of the Dice loss is that is is prone to optimization instability, resulting from gradient calculations involving small denominators Mukhoti, Kulharia, Sanyal, Golodetz, Torr & Dokania (2020).

Once one or a combination of the losses above $\mathcal{L}$ is chosen, the optimization model for supervised segmentation takes the following general form:

$$\min_{\boldsymbol{\theta}} \quad \sum_{i \in \Omega} \mathcal{L}(y(i), p(i, \boldsymbol{\theta})) \tag{1.26}$$

1.3.2.2 CRF as post-processing

A caveat of CNN segmentation models is that they perform pixel-level classification and do not model neighborhood dependencies directly. To address this problem, the popular segmentation network U-net Ronneberger, Fischer & Brox (2015) (see Section 1.3.5) combines both low-level and high-level features, but there is still no guarantee of spatial consistency in the final segmentation map. Empirically, it is seen that CNN are able to discover some neighborhood dependencies, leading to relatively smooth segmentation maps. Nonetheless, holes and isolated regions remain frequent. Moreover, contrary to other ML methods, CNN generally do not explicitly model shape and edge constraints. As a result, the final segmentation can appear rougher.

Typically, conditional random fields (CRF) Blake, Kohli & Rother (2011) are adopted to produce a much more refined segmentation output, as CRF can directly model spatial structures (dependencies between regions, shape, region connectivity, etc.). The most popular approach is DeepLab, a fully connected Gaussian CRF model proposed in Chen *et al.* (2018b) where the unary potentials are supplied by a CNN. The key attractive component is its approximate but fast CRF solving method, first introduced in Krähenbühl & Koltun (2011). Figure 1.1 presents an overview of DeepLab.

While the supervised methods presented above lead to state-of-the-art performance given a large enough training dataset, they require its pixel-level annotation, a cumbersome task. In parallel, efforts have been made to develop weakly and semi-supervised methods. Instead of requiring

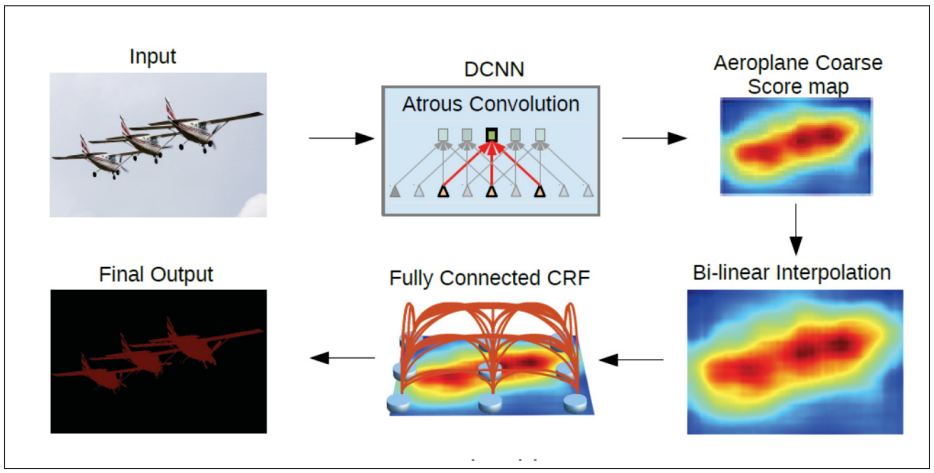


Figure 1.1 Deeplab framework to refine output segmentation masks. Image source: Chen *et al.* (2018b)

detailed annotations, these methods build predictive models using partial annotations, which are less costly to obtain. We will briefly present them below.

1.3.3 Semi or weakly supervised segmentation losses

1.3.3.1 Partial annotations

Instead of requiring detailed but time-consuming annotations, some faster (though imperfect) alternatives can be envisioned, as illustrated in Figure 1.2. They can be regrouped in two broad categories: *sparse* (or semi) annotations and *weak* annotations. Let us denote $\Omega_L \subseteq \Omega$ the set of labeled pixels, and $\Omega_U \subseteq \Omega$ the set of unlabeled pixels, such as $\Omega_L \cup \Omega_U = \Omega$ and $\Omega_L \cap \Omega_U = \{\emptyset\}$.

Sparse annotations: These annotations provide known class values for a subset Ω_L of pixels. Examples of such annotations include scribbles and point annotations. The class value of unlabeled pixels Ω_U is unknown. $\Omega_L = \Omega$ corresponds to the fully annotated setting.

Weak annotations: These annotations provide image-level or region-level information on the image. However, uncertainty with regard to pixel-level class values remains. Example of such

annotations are image-level labels, or image-level tags, where we are given the information of the classes found in the whole image (see Figure 1.2).

Other examples are bounding-box annotations: no pixels outside the bounding box belong to the object, but *some* pixels inside do, although we have no information regarding which ones. Other forms of weak labels may include higher-level information, such as the size of anatomical structures, or other information regarding the global shape, such as radius, etc.

Note that this type of weak annotations can be derived from radiological reports, or from user interactions, making them an interesting setting to consider. However, when designing a method that uses weak labels, the uncertainty regarding these annotations should be taken into account.

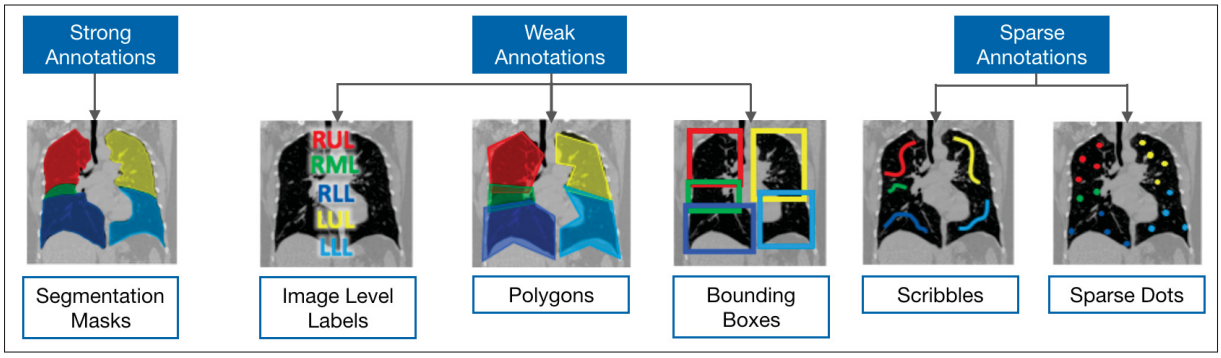


Figure 1.2 Comparing strong and weak annotations for lung lobes of a chest CT scan. Image adapted from Tajbakhsh *et al.* (2020)

1.3.3.2 Self-training

One way to approach the lack of annotations is to consider the unknown labels as hidden variables that can be learned. In the case of segmentation tasks, pseudo-labels of images are generated via the network predictions. Such methods fall in the self-training paradigm, and take the following general form:

$$\min_{\boldsymbol{\theta},\tilde{y}} \quad \sum_{i\in\Omega} \mathcal{L}(\tilde{y}(i), p(i,\boldsymbol{\theta})) \tag{1.27}$$

where $\tilde{y}$ are *pseudo-labels* or *proposals* on target predictions. Such an assumption allows to choose for $\mathcal{L}$ standard segmentation losses from Section 1.3.2, such as cross-entropy. By

minimizing the loss above with respect to $\tilde{y}$, the optimized pseudo-labels should approximate the underlying true target ground truth.

Self-training methods generally alternate between generating the pseudo-labels, and then training with them, and can be seen as a form of expectation maximization. Common strategies often involve complex heuristics and scheduling procedures for choosing thresholds Sohn, Berthelot, Carlini, Zhang, Zhang, Raffel et al. (2020); Xie, Luong, Hovy & Le (2020).

The main difficulty is that some errors will occur in the generated pseudo-labels. Therefore, self-training methods are inherently unstable, due to error accumulations, caused by the mispredicted pseudo-labels which can reinforce themselves by training the network with wrong information Liang, He, Sun & Tan (2019). For instance, many medical imaging applications suffer from data imbalance, with orders of magnitude more background pixels than foreground ones. In these cases, using an unmodified self-training loss can produce a collapse to a trivial solution, i.e. assign all probabilities to the most probable class, which is the background Zou, Yu, Kumar & Wang (2018). An example of this collapse can be seen in Figure 1.3.

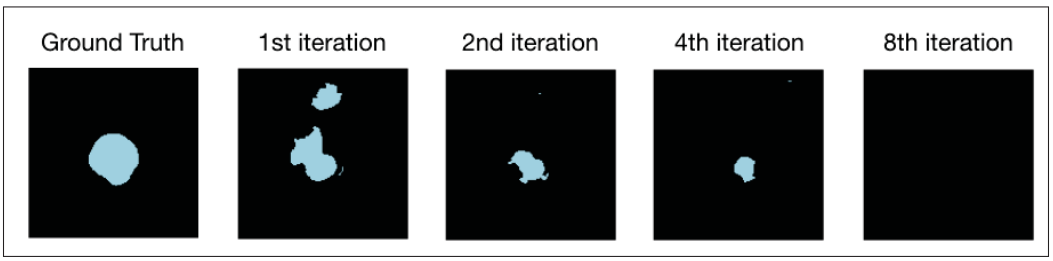


Figure 1.3 Example of a collapse of the predictions using self-training proposal loss on prostate segmentation. The predicted segmentation mask gradually disappears, the model eventually classifies every pixel to the dominant background class

1.3.3.3 Optimization-based image segmentation

Semi or weakly supervised segmentation methods are often formulated as energy minimization problems Nosrati & Hamarneh (2016). An energy function is typically divided into two main terms, the data term(s), and the regularization term(s), similarly to Eq.(1.5) from Section

(1.1.2). The data term expresses how strongly a pixel should correspond to a specific label. The regularization terms represents a preference for specific solutions. For instance, the regularization can be a prior restricting the search space to anatomically plausible solutions and penalizing any deviation from such prior. For a given image I and its softmax output p_{θ}, the energy minimization is written as :

$$E(p_{\theta}) = \sum_{m} D_m(p_{\theta}; I) + \lambda \sum_{j} R_j(p_{\theta}) \tag{1.28}$$

where D_m are the data terms and R_j are the regularization terms.

As highlighted by Nosrati & Hamarneh (2016); Tang, Perazzi, Djelouah, Ben Ayed, Schroers & Boykov (2018b), such a formulation allows to 1) directly incorporate multiple competing criteria in a differentiable end-to-end approach, 2) quantitatively measure the extent by which a method satisfies the different terms, and 3) examine the relative performance of different solutions.

However, there is a trade-off between fidelity and optimization when choosing the energy minimization function. Fidelity describes how faithful the energy function is to the data and how accurately it can model and capture the intricate details of the problem. Oftentimes though, a more faithful objective function model leads to a more complex optimization problem. Conversely, when foregoing fidelity to obtain an objective function that is easier to optimize, the solution may not be precise enough for the segmentation target. We detail below an example of optimization-based image segmentation with scribbles.

1.3.3.4 Semi-supervision with scribbles for in-domain segmentation networks

How to best employ the partial annotations on the image domain Ω is central to semi-supervised approaches. One way is to consider the unknown labels as hidden variables that can be extended from the partial annotations. In the case of segmentation tasks, pseudo-labels $\tilde{y}$ of images can be generated via the network predictions and the partial annotations.

When scribbles are available, a common method is to apply CRF with the CNN outputs being used as unary weights Tang, Djelouah, Perazzi, Boykov & Schroers (2018a); Zheng, Jayasumana, Romera-Paredes, Vineet, Su, Du et al. (2015). The optimisation typically alternates between generating segmentation pseudo-labels $\tilde{y}$ extending from the scribbles (e.g. via a mean-field approach Krähenbühl & Koltun (2011)) and training the CNN with supervision to mimic these pseudo-labels via a self-training loss :

$$\min_{\boldsymbol{\theta},\tilde{y}} \sum_{i\in\Omega} \mathcal{L}(\tilde{y}(i), p(i,\boldsymbol{\theta})) \tag{1.29}$$

In Tang *et al.* (2018b), Tang et al. showed that this is in fact an approximation of an end-to-end approach using a direct loss such as Eq.(1.28):

$$\sum_{i\in\Omega_L} \mathcal{L}(y(i), p(i,\boldsymbol{\theta})) + \lambda\mathcal{R}(p(i,\boldsymbol{\theta})) \tag{1.30}$$

where Ω_L is the set of labeled pixels (the scribbles), $\mathcal{R}$ is a regularizer function and λ a scalar balancing the two objectives.

A common choice for regularization term R is an energy function of pairwise potentials, encouraging spatial coherence by penalizing discontinuities between neighbouring pixels of image I. A quadratic relaxation of the Potts model is often used : $E(I) = \sum_{i,j\in\Omega} W_{ij}\,[p(i,\boldsymbol{\theta}) \neq p(j,\boldsymbol{\theta})] \approx \sum_{i,j\in\Omega} W_{ij}\,\|p(i,\boldsymbol{\theta}) - p(j,\boldsymbol{\theta})\|^2$. where $W = W_{ij}$ is a matrix of pairwise discontinuity costs or an affinity matrix. The key step is the relaxation on the right hand side of the equation above with $p(i,\boldsymbol{\theta}) \in [0,1]^K$, allowing for the use of the softmax output of the CNN, instead of traditional binary class indicators $p(i,\boldsymbol{\theta}) \in \{0,1\}^K$. Therefore this approximation yields a fully differentiable $E(I)$ which can be integrated in a direct loss minimization scheme such as Eq. (1.30).

In the following Section, we present another important tool for semi, weakly, or unsupervised segmentation, i.e. constrained deep networks, which we will integrate in our DA methods. We first briefly review constrained optimization for convex problems. We then present a few recent

works which have proposed constrained deep networks to guide semi- or weakly- supervised learning.

1.3.4 Constrained deep networks

1.3.4.1 Overview of constrained optimization methods

Let us reconsider the optimization problem from Section 1.1.1. Often, we want to minimize a function F while also enforcing some conditions on the solution; those are constraints on the optimization process, which the solution should satisfy. Formally, constraints can be written as follows[3]:

$$\begin{aligned} \min_{x \in \mathbb{R}^D} \quad & F(x) \\ \text{subject to} \quad & f_1(x) \leq 0 \\ & \dots \\ & f_C(x) \leq 0. \end{aligned} \tag{1.31}$$

The *feasible region* C, or feasible set, or search space is defined as the set of all possible x that satisfy the problem constraints : $C = \left\{x \in \mathbb{R}^D \mid f_1(x) \leq 0, f_2(x) \leq 0, ..., f_C(x) \leq 0\right\}$. The problem in Eq. (1.31) is equivalent to :

$$\min_{x \in C} \quad F(x) \tag{1.32}$$

[3] An equality constraint $f_n(x) = 0$ can be written with two inequality constraints: $f_i(x) \leq 0$ and $-f_i(x) \leq 0$.

1.3.4.1.1 Quadratic Penalty

One of the earliest and simplest methods, the naive quadratic penalty (QP) method Bertsekas (1976) converts the constrained optimization problem in Eq. (1.31) to an unconstrained optimization problem by adding the constraints to the objective function as a quadratic penalty term:

$$\min_{x \in \mathbb{R}^D} F(x) + \mu \sum_{i=1}^{C} \|f_i(x)\|^2 \tag{1.33}$$

where the penalty parameter μ is a positive scalar, which is not optimized by QP. To ensure convexity, convergence, and satisfaction of the constraints, μ should be increased indefinitely over the iterations, which leads to training instability. Another major drawback of QP is that there is no guarantee that the solution will be in the feasible set C.

1.3.4.1.2 Projected Gradient Descent

Projected Gradient Descent (PGD) is a standard way to solve the constrained optimization problem in Eq. (1.31), by modifying the update of GD Algorithm (1.1) by :

$x^{t+1} \leftarrow P_C(x^t - \tau \nabla F(x^t))$ where $P_C(.)$ is a projection operator, and itself is also an optimization problem:

$$P_C(x_0) = \arg\min_{x \in C} \frac{1}{2} \|x - x_0\|_2^2 \tag{1.34}$$

The idea of PGD is simple: if the point $x^t - \tau \nabla F(x^t)$ after the gradient update is leaving the set C, project it back. PGD is an "economic" algorithm if the problem is easy to solve. This is not true for general C, when there are many constraint sets that are difficult to project onto. Fortunately, the Lagrangian method which we will see below is more general and is typically used in these complex cases.

1.3.4.1.3 Method of Lagrangian multipliers

The fundamental algorithm for constrained optimization was introduced by Lagrange, through the Lagrangian-dual problem described below. We denote an optimal solution f^* and its corresponding optimal input x^*. Let us rewrite Eq. (1.31) as an unconstrained optimization problem, exploiting an infinite penalty function when the constraints are not satisfied:

$$\min_x F(x) + \sum_{i=1}^{C} \infty_{[f_i(x)>0]}, \tag{1.35}$$

where $\infty_{[a]}$ takes the value 0 when a is `False`, and the value $+\infty$ when a is `True`. The function above is highly discontinuous, and therefore very difficult to optimize. However notice that:

$$\forall i : \infty_{[f_i(x)>0]} = \max_{\lambda \geq 0} \lambda_i f_i(x). \tag{1.36}$$

When maximizing over $\boldsymbol{\lambda} \geq 0$ for some $f_i(x)$, the optimal solution is 0 when $f_i(x) < 0$. When $f_i(x)$ is positive, the optimal $\boldsymbol{\lambda}$ value is $+\infty$. The unconstrained minimization problem Eq. (1.35) can therefore be rewritten as:

$$\min_x \max_{\lambda \geq 0} \quad F(x) + \sum_{i=1}^{C} \lambda_i f_i(x). \tag{1.37}$$

This problem is still difficult to optimize. However, consider swapping the minimization and maximization as follows:

$$\max_{\lambda \geq 0} \min_x \quad F(x) + \sum_{i=1}^{C} \lambda_i f_i(x). \tag{1.38}$$

While easier to solve, this problem does not necessarily yield the same optimum as the original Eq. (1.37).

The Lagrange function is defined as : $\mathcal{L}(x, \boldsymbol{\lambda}) = F(x) + \sum_{i=1}^{C} \lambda_i f_i(x)$, where the λ_i are called the Lagrange multipliers.

For a fixed $\boldsymbol{\lambda}$, we can optimize x, this is the Lagrangian dual function:

$$\mathcal{L}_\lambda(\boldsymbol{\lambda}) = \min_x F(x) + \sum_{i=1}^{C} \lambda_i f_i(x). \tag{1.39}$$

We can easily show that $\mathcal{L}_\lambda(\boldsymbol{\lambda}) \leq f^*$. Indeed, for a feasible solution $\tilde{x}$, $\forall i \in \{1, ..., P\}$: $\lambda_i f_i(\tilde{x}) \leq 0$. Therefore,

$$\mathcal{L}_\lambda(\boldsymbol{\lambda}) \leq F(\tilde{x}) + \sum_{i=1}^{C} \lambda_i f_i(\tilde{x}) \leq F(\tilde{x}) \leq F(x^*) = f^*. \tag{1.40}$$

The difference $f^* - \mathcal{L}_\lambda(\boldsymbol{\lambda})$ is called the *duality gap*. By alternating optimization with respect to $\boldsymbol{\lambda}$ and x, the duality gap is decreased. If the *Karush-Kuhn-Tucker* (KKT) conditions are met Boyd & Vandenberghe (2004), strong-duality holds and the duality gap is zero : $\hat{x} = x^*$. This is generally not the case, even for convex settings.

Still, we can summarize the following properties observed for convex problems, contrary to Quadratic Penalty methods, making Lagrangian optimization a central tool in optimization Boyd & Vandenberghe (2004): first, the optimal weights of the constraints λ_i can be obtained automatically; second, the Lagrangian acts as a barrier for satisfied constraints; and third, it provides a guarantee that the constraints will be satisfied when feasible solutions exist.

1.3.4.1.4 Interior Point methods

Interior point techniques such as the log-barrier methods Boyd & Vandenberghe (2004) are popular alternatives to Lagrangian optimization as they may avoid its costly dual-updates while still offering nice convergence and optimality properties. Interior-point approaches require a *feasible starting point*, where all constraints are satisfied. The constrained optimization problem in Eq. 1.31 can then be optimized using an additional barrier that tends to infinity when the

constraints approach their upper bound (see Figure 1.4). The new optimization problem is:

$$\min_{x \in \mathbb{R}^D} F(x) + \sum_{i=1}^{C} \psi_t(f_i(x)) \qquad \psi_t(z) = -\frac{1}{t} log(-z) \tag{1.41}$$

where $t > 0$ increases over time.

The main difficulty regarding interior point methods such as log-barrier methods is that finding a feasible starting point may not be possible analytically, depending on the optimization problem.

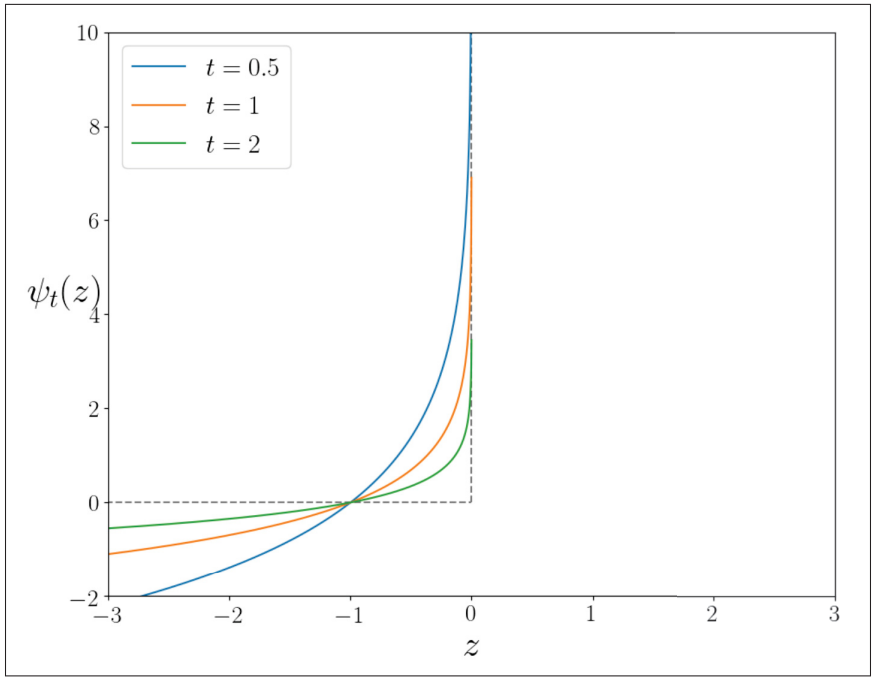


Figure 1.4 Parameterized log-barrier, for different t values

1.3.4.1.5 Augmented Lagrangian methods

In Augmented Lagrangian methods (ALM), the original problem is transformed into an unconstrained optimization problem, by combining the penalty concept and the primal-dual idea of the Lagrangian function. Specifically, instead of adding the penalty term to the objective

function, it is added to the Lagrangian function, creating the Augmented Lagrangian Function:

$$\mathcal{L}_{\lambda,\mu}(x,\boldsymbol{\lambda}) = F(x) + \mu \sum_{i=1}^{C} \|f_i(x)\|^2 + \sum_{i=1}^{C} \lambda_i f_i(x) \tag{1.42}$$

In practice, this is solved by the min-max optimization problem:

$$\max_{\boldsymbol{\lambda}\geq 0} \min_{x} \mathcal{L}_{\lambda,\mu}(x,\boldsymbol{\lambda}). \tag{1.43}$$

with the following updates: $\boldsymbol{\lambda}_i^{t+1} = \boldsymbol{\lambda}_i^t + \mu f_i(x)$, and $\left\{\mu^t\right\}$ is a positive increasing sequence.

Unlike the quadratic penalty method, increasing the $\left\{\mu^t\right\}$ indefinitely is not required in the ALM, alleviating training instability. Additionally, contrary to the method of Lagrangian multipliers, the ALM does not need a convexity assumption to guarantee convergence.

Next, we discuss the integration of constrained optimization with deep neural networks.

1.3.4.2 Lagrangian dual optimization with CNN

Let us reformulate the standard method for solving constrained optimization in Equation (1.31) to the CNN setting, with the notation from Section 1.3.2:

$$\begin{aligned} \min_{\boldsymbol{\theta}} \quad & \sum_{n=1}^{N} \mathcal{L}(\mathbf{y}_n, \mathbf{p}_n(\boldsymbol{\theta})) \\ \text{subject to} \quad & f_{1,1}(\mathbf{p}_1(\boldsymbol{\theta})) \qquad\qquad \ldots \\ & f_{C,1}(\mathbf{p}_1(\boldsymbol{\theta})) \\ & \ldots \\ & f_{C,N}(\mathbf{p}_N(\boldsymbol{\theta})) \end{aligned} \tag{1.44}$$

where $(\mathbf{y}_n)_{n\leq N}$ are partially or completely unknown, and C is the maximum number of constraints to enforce on every single sample n, which are denoted $f_{1,n}...f_{C,n}$. The corresponding Lagrangian is:

$$\max_{\boldsymbol{\lambda}\geq 0}\min_{\boldsymbol{\theta}} \quad \sum_{n=1}^{N}\mathcal{L}(\mathbf{y}_n,\mathbf{p}_n(\boldsymbol{\theta})) + \sum_{c=1}^{C}\sum_{n=1}^{N}\lambda_{c,n}f_{c,n}(\mathbf{p}_n(\boldsymbol{\theta})) \tag{1.45}$$

where $\boldsymbol{\lambda}\in\mathbb{R}_{+}^{C\times N}$ is the dual variable (or Lagrange-multiplier) vector, with $\lambda_{c,n}$ the multiplier associated with constraint $f_{c,n}(\mathbf{p}_n(\boldsymbol{\theta})\leq 0$.

Extending constrained optimization to a CNN setting is far from straightforward. Even if possible, the convergence and optimality guarantees of Lagrangian optimization are lost in this highly non-convex problem. The main difficulty is that solving Eq. 1.45 with Lagrangian optimization requires alternating between the two following steps: (i) the CNN network is trained using backpropagation with respect to the network parameters $\boldsymbol{\theta}$; (ii) the Lagrange multipliers are updated following a standard projected gradient ascent, updating the dual variable $\boldsymbol{\lambda}$. As training a neural network can take from a few hours to several days, retraining the neural network *at each iteration* is prohibitive. Moreover, these two optimization steps may be ill-fitted: alternating these them may lead to oscillation and cancel the benefits of stochastic optimization. This detrimental interplay was experimentally verified in Márquez-Neila et al. (2017), whereby replacing the hard constraints with simple soft quadratic penalties yielded better results than Lagrangian optimization.

Despite the clear benefits of imposing hard constraints on CNN in many applications, the limitations above deter the use of standard Lagrangian-dual optimization. In the literature, alternative ways of incorporating constrained formulations for deep networks have been proposed. For instance, in Sangalli, Erdil, Hötker, Donati & Konukoglu (2021), the training of a deep network for binary classification with under-represented classes is posed as a constrained optimization problem, and solved using augmented Lagrangian methods.

We present below three recent contributions incorporating a constrained formulation for deep segmentation networks. In these works, prior knowledge about the segmentation regions is embedded in the CNN optimization problem, thereby compensating for the lack of annotations.

1.3.4.3 Constraining deep segmentation networks via Lagrangian with proposals

In Pathak, Krähenbühl & Darrell (2015), Pathak et al. embed deep CNN with linear constraints in a weakly supervised segmentation problem. They investigate whether the optimization of segmentation networks can be guided by imposing constraints on the sizes of predicted structures. Their original formulation is the following :

$$\begin{aligned} \min_{\boldsymbol{\theta}} \quad & \sum_{n} \mathcal{L}(\mathbf{y}_n, \mathbf{p}_n(\boldsymbol{\theta})) & \\ \text{s.t.} \quad & \mathbf{p}_n(\boldsymbol{\theta})^\top a_1 - b_1 \leq 0 & 1 \leq n \leq N \\ & \ldots & \\ & \mathbf{p}_n(\boldsymbol{\theta})^\top a_C - b_C \leq 0 & 1 \leq n \leq N. \end{aligned} \tag{1.46}$$

where $\mathbf{y}_n$ is partially or completely unknown, $a_1, ..., a_C \in \mathbb{R}^{|\Omega|}$ and $b_1, ..., b_C \in \mathbb{R}^K$.

Instead of the intractable Lagrangian formulation, they introduce an approximate Lagrangian optimization which uses the constraints to synthesize pseudo-labels as training masks. This allows to mimic full supervision, while avoiding the dual optimization. Specifically, y_n is converted to a continuous variable ($y_n \in [0,1]^{K \times |\Omega|}$), and a latent variable $\tilde{y}^n \in [0,1]^{K \times |\Omega|}$ serves as a pseudo-label. Thus rather than imposing the linear constraints on the network output,

they are imposed on the pseudo-labels in the following way:

$$\begin{aligned} \min_{\tilde{y},\boldsymbol{\theta}} \quad & \sum_{n} \mathrm{KL}(\tilde{y}_n || p_n(\boldsymbol{\theta})) && (1.47)\\ s.t. \quad & \tilde{y}_n^\top a_1 - b_1 \preceq 0 && \forall n \leq N \\ & \dots \\ & \tilde{y}_n^\top a_C - b_C \preceq 0 && \forall n \leq N \\ & \mathbf{1}^\top \tilde{y}_n(i) = 1 && \forall n \leq N, \forall i \in |\Omega|. \end{aligned}$$

where KL denotes the Kullback–Leibler divergence[4], pushing the CNN softmax output to match the latent variable $\tilde{y}$.

The corresponding Lagrangian is:

$$\begin{aligned} \max_{\boldsymbol{\lambda},\boldsymbol{\nu}} \min_{\tilde{y},\boldsymbol{\theta}} \quad & \sum_{n} \left(\mathrm{KL}(\tilde{y}_n || p_{\boldsymbol{\theta}}(n)) + \sum_{i=1}^{C} \lambda_{i,n} (\tilde{y}_n^\top a_i - b_i) + \sum_{i \in \Omega} \nu_{i,n} (\mathbf{1}^\top \tilde{y}_n(i) - 1) \right) && (1.48)\\ s.t. \quad & \boldsymbol{\lambda} \geq 0 \end{aligned}$$

where $\boldsymbol{\lambda} \in \mathbb{R}_+^{C \times N}$ and $\boldsymbol{\nu} \in \mathbb{R}^{N \times |\Omega|}$ are the Lagrangian dual variables. Minimizing $\tilde{y}$, for constant $\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\nu}$, can be solved analytically. Updating $\boldsymbol{\lambda}$ and $\boldsymbol{\nu}$ requires performing a projected gradient ascent.

Therefore, the stochastic gradient descent learning of the network parameters is decoupled from the constrained optimization, i.e. the optimizing w.r.t the latent variable, which corresponds to proposal generation subject to the constraints. The introduction of latent variable $\tilde{y}_n$ makes Pathak *et al.* (2015) a proposal-based method with the drawbacks described in Section

[4] For discrete probability distributions P and Q defined on the same sample space, $\mathcal{X}$, the Kullback–Leibler divergence from Q and P is defined to be $D_{\mathrm{KL}}(P \| Q) = \sum_{x \in \mathcal{X}} P(x) \log \left(\frac{P(x)}{Q(x)} \right)$.

1.3.3.2. Specifically, initial training errors might reinforce themselves or cause the training to become unstable when only incomplete labels (like scribbles) are available. Additionally, the computational complexity of the two-step procedure can be prohibitive.

1.3.4.4 Constraining deep segmentation networks via Naive Penalty

In Kervadec, Dolz, Tang, Granger, Boykov & Ben Ayed (2019b), Kervadec et al. build on the work by Pathak et al Pathak *et al.* (2015), also embedding a constraint on the size of the predicted structures. However, contrary to Pathak *et al.* (2015), they do away with dual optimization, introducing the constraint in the form of a naive penalty term in the loss function instead. This penalty term is differentiable and fully suited for standard stochastic gradient descent. This transforms the hard constraints of Eq. (1.44) in soft ones.

For each slice n, the predicted size of the segmentation structures $k = 1...K$ in n is estimated as : $\mathbf{V}_n = \sum_{i \in \Omega} \mathbf{p}_n(i, \boldsymbol{\theta}) \in \mathbb{R}^K$. The inequality constraint to integrate is: $\mathbf{a} \leq \mathbf{V}_n \leq \mathbf{b}$, where $(\mathbf{a},\mathbf{b}) \in \mathbb{R}_+^K \times \mathbb{R}_+^K$.

The formulation in Eq. (1.47) is transformed into a fully differentiable problem :

$$\min_{\boldsymbol{\theta}} \quad \sum_n \mathcal{L}(\mathbf{y}_n, \mathbf{p}_n(\boldsymbol{\theta})) + \lambda C\left(\mathbf{V}_n\right) \tag{1.49}$$

where they choose C:

$$C\left(\mathbf{V}_n\right) = \begin{cases} \left(\mathbf{V}_n - \mathbf{a}\right)^2, & \text{if } \mathbf{V}_n < \mathbf{a} \\ \left(\mathbf{V}_n - \mathbf{b}\right)^2, & \text{if } \mathbf{V}_n > \mathbf{b} \\ 0, & \text{otherwise} \end{cases} \tag{1.50}$$

Note that by embedding the constraints through a naive penalty, there is no guarantee that the predictions obtained will satisfy these constraints. However, surprisingly, the authors observe experimentally better results than the Lagrangian-based constrained CNN in Pathak *et al.* (2015), while greatly reducing the computational cost. Figure 1.5 shows the improvement obtained

by Kervadec *et al.* (2019b) for a semi-supervised binary segmentation problem (left ventricle segmentation with only 1% of labeled pixels), exploiting a constraint on the size of the ventricle.

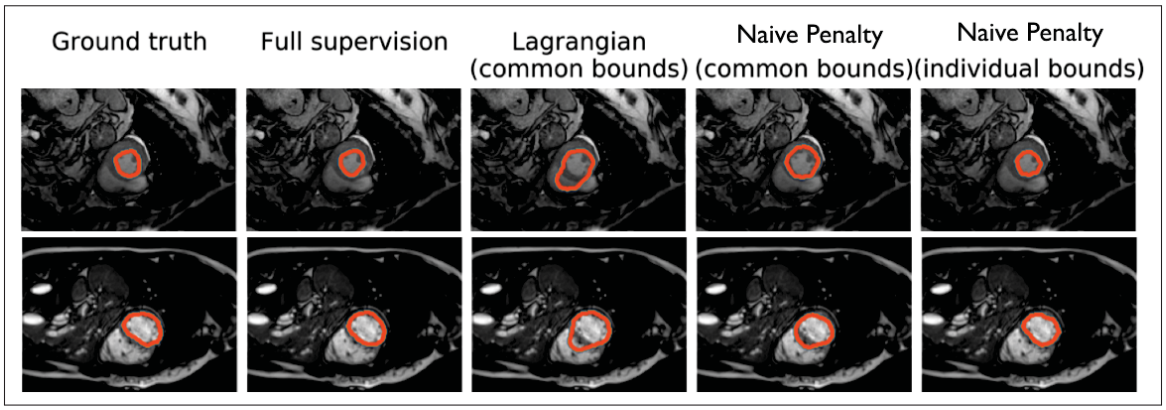


Figure 1.5 Visual comparison of different methods studied in Kervadec *et al.* (2019b), showing the superiority of Naive Penalty over the Lagrangian-based constrained CNN in Pathak *et al.* (2015). Figure adapted from Kervadec *et al.* (2019b)

Amongst the limitations of this work, we should mention that the size constraints are supervised: for each slice n, they are obtained from its ground truth mask, form which a, b are derived by adding margins of error. This is not a realistic scenario in practice. For greater applicability, these constraints should be learned or estimated without the use of ground truth masks.

1.3.4.5 Constraining deep segmentation networks Log-barrier methods

The simplicity of penalty methods such as the one in Kervadec *et al.* (2019b) comes at a price. First, there is no guarantee that the predictions obtained will satisfy the constraints. Second, if multiple constraints are desired, the relative importance of each penalty term in the overall function requires careful tuning. More importantly, in the case of several competing constraints, penalties do not act as barriers at the boundary of the feasible set. This is because a satisfied constraint yields a null penalty and null gradient. As a result, a subset of constraints that are satisfied at one iteration may not be satisfied at the next. This oscillation could impede the

optimization from converging to a solution satisfying all constraints Kervadec, Dolz, Yuan, Desrosiers, Granger & Ben Ayed (2022).

To address these issues, Kervadec et al. propose a log-barrier method in Kervadec *et al.* (2022) to approximate Eq. (1.31). Their method also avoids the iterative Lagrangian dual / CNN optimization. The constraints are directly introduced into stochastic optimization, with implicit dual variables, via a log barrier function. Instead of the classical log barrier function in Eq. 1.41, they propose an extension :

$$\tilde{\psi}_t(z) = \begin{cases} -\frac{1}{t}\log(-z) \text{ if } z \leq -\frac{1}{t^2} \\ tz - \frac{1}{t}\log\left(\frac{1}{t^2}\right) + \frac{1}{t} \qquad \text{otherwise} \end{cases} \tag{1.51}$$

where contrary to ψ_t, the domain of $\tilde{\psi}_t$ is not restricted to feasible points θ, greatly simplifying the initialization of θ. The gradient of $\tilde{\psi}$ is strictly positive, and increases when a satisfied constraint approaches violation during optimization, pushing it back towards feasible points, thereby acting as a barrier.

Therefore, the following unconstrained log-barrier extension is optimized:

$$\min_{\boldsymbol{\theta}} \sum_{n=1}^{N} L(\mathbf{y}_n, \mathbf{p}_n(\boldsymbol{\theta})) + \sum_{c=1}^{C}\sum_{n=1}^{N} \tilde{\psi}_t\left(f_{c,n}\left(\mathbf{p}_n(\theta)\right)\right) \tag{1.52}$$

When $t \to +\infty$, this extension approaches the initial problem with hard constraints in Eq. (1.31). In practice, t is progressively augmented at each iteration of the optimization procedure.

Figure 1.6 shows the improvement obtained in Kervadec *et al.* (2022) on a toy dataset, integrating constraints on the size and the centroid of the circles via their log-barrier method, compared to the Lagrangian method in Pathak *et al.* (2015), and to the naive quadratic penalty in Kervadec *et al.* (2019b).

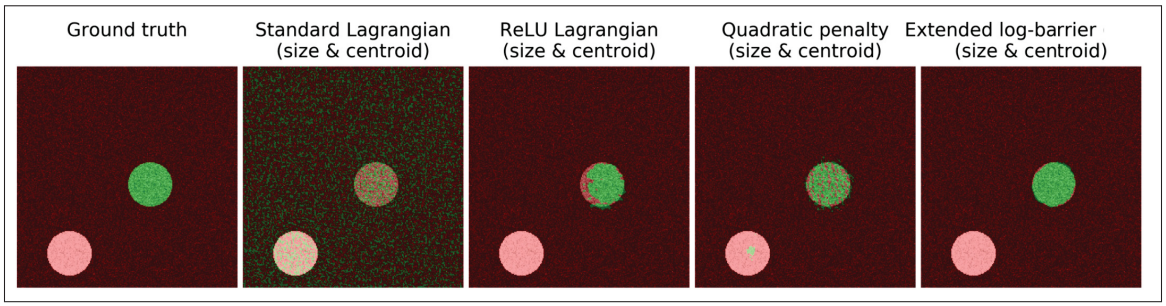


Figure 1.6 Visual comparison of different methods studied in Kervadec *et al.* (2022), showing the superiority of log-barrier constrained CNN on a toy example. Figure courtesy to Kervadec *et al.* (2019b)

After introducing supervised, semi and weakly-supervised segmentation, as well as constrained deep networks, the following Section summarizes the most popular deep-learning architectures used in medical image segmentation.

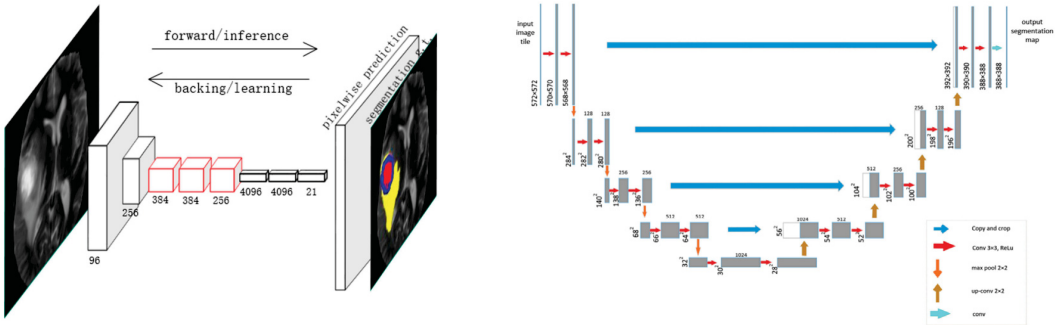
1.3.5 The state-of-the-art of segmentation architectures

The pioneering work for deep semantic segmentation was introduced in Long, Shelhamer & Darrell (2015a), in the form of a fully convolutional network (FCN) able to produce pixel-level predictions of the original image size. The fully connected layers standard in networks such as the popular VGG are replaced by additional convolutional layers, a deconvolution layer to upsample the feature map of the last convolution layer and restore it to the same size of the input image. A softmax layer is added to obtain the pixel classification. The FCN network structure is shown in Figure 1.7 (left).

The DeepLab model Chen, Papandreou, Kokkinos, Murphy & Yuille (2015) addressed the shortcomings of the FCN by using atrous convolutions on pretrained CNN models ResNet-101/VGG-16 He, Zhang, Ren & Sun (2016). Compared with traditional convolutions, these atrous convolutions allow to expand the receptive field and to increase the density of features without increasing the number of parameters. Finally, they introduced conditional random field (CRF) Krähenbühl & Koltun (2011) to produce fine segmented output, which has since

them become a standard post-processing step. Improvements over the DeepLab model have been proposed including DeepLabv2 Chen *et al.* (2018b), DeepLabv3 Chen, Papandreou, Schroff & Adam (2017a).

Also derived from FCN, Ronneberger et al. designed the UNet in Ronneberger *et al.* (2015), one of the most widely used networks for both medical and natural image segmentation. The UNet is a 19 layer deep network with an encoding path and a decoding path using deconvolution layers, forming a U channel. UNet also includes the use of long skip connections between the layers of equal resolution in the encoding path and decoding path. This is the key idea to preserve the high-resolution spatial information which lacks in the initial FCN. Therefore the strength of UNet relies on its combination of low-level and high-level information. The low-level information helps to improve accuracy, while the high-level information helps to extract complex features. The UNet network structure is shown in Figure 1.7 (right).



a) The structure of the fully convolutional network (FCN) Long *et al.* (2015a).

b) The structure of the UNet Ronneberger *et al.* (2015).

Figure 1.7 Illustration of different modern segmentation networks

Many variants have been developed on the UNet, adding new modules or integrating other design concepts. A popular one is the V-Net proposed by Milletari et al. in Milletari *et al.* (2016). Unlike UNet, V-Net involves residual blocks He *et al.* (2016) as short skip connections between early and later convolutional layers, and uses Dice loss instead of the standard binary cross entropy loss to cope with the class imbalance problem. Another important variant is the

3D UNet model introduced in Çiçek, Abdulkadir, Lienkamp, Brox & Ronneberger (2016), with a similar network structure as the UNet, but extended to perform 3D image segmentation.

The aforementioned segmentation models have achieved state-of-art performance in many medical image segmentation tasks. The choice of a particular deep learning model depends on factors such as the body part to be segmented, the imaging modality employed, and the type of disease which each call for different requirements. For instance, the segmentation area for the whole heart, the lungs, the brain white matter is relatively large, while the segmentation of blood vessels in retinal images needs specialized losses and techniques. Segmentation of lesions combines the challenges of object detection and substructure segmentation; therefore segmenting tumours, such as brain, skin and breast tumours, or lung nodules, requires careful design, as little a-priori anatomical knowledge can be drawn from.

State-of-the-art results have been obtained by supervised deep segmentation models for the **brain:** CNN segmentation methods were the top rankings techniques in successive brain tumor segmentation challenges (BRATS, ISLES, MRBRains) Ghafoorian, Karssemeijer, Heskes, van Uder, de Leeuw, Marchiori et al. (2016); **heart:** CT deep segmentation methods have been FDA-cleared for coronary artery visualization HeartFlowNXT (2017); the current state-of-the-art for deep cardiac image segmentation is summarized in Chen, Qin, Qiu, Tarroni, Duan, Bai et al. (2020a); **prostate:** a multi-scale segmentation network from Jia, Cai, Huang & Xia (2022) is the top performer in the PROMISE12 prostate segmentation challenge; **liver:** Czipczer & Manno-Kovacs (2022) modified the 3D U-Net to obtain state-of-the-art results on three public liver segmentation datasets; **lesion segmentation:** Mirikharaji & Hamarneh (2018) proposed a state-of-the-art fully convolutional network framework and won the ISBI 2017 skin segmentation challenge.

To summarize, CNN segmentation models have been developed for most conditions for which data can be collected. They are increasingly achieving or surpassing human-level performance in many medical imaging applications when a large set of in-domain annotated images are available. However, these controlled settings are not realistic in clinical applications. Additionally,

improvements in accuracy delineation metrics are not always synonymous with practical improvement, where robustness may be more important. Hence the medical computation community has turned its focus to learning models which will be applicable in wider settings, such as models robust to domain shifts, which is the focus of the next section.

1.4 Domain adaptation for medical image segmentation

Deep learning models typically assume that the training dataset (source domain S) and test dataset (target domain T) share the same data distribution. However, in medical applications, distribution differences between training and test datasets occur frequently. This leads to poor performance in the target dataset, and is referred to as the "domain shift" problem, impeding deployment in practice. Although labeling each new test dataset could close the performance gap, this is generally expensive, requiring labor-intensive participation of radiologists.

Domain adaptation (DA) is a growing field devoted to tackling this performance gap, with no additional annotations (unsupervised domain adaptation, UDA) or minimal annotations (weakly-supervised or semi-supervised domain adaptation) in the target domain. In DA, the source domain and target domain share the same learning task. In the following sections, we first describe the main sources of domain shifts in medical imaging. We then introduce the problem which we are trying to solve, and describe desirable properties of a DA framework. Subsequently, we present a major branch of DA methods, i.e. adversarial methods, and highlight their limitations, most importantly their unrealistic setting and cumbersome framework. We finally present a few key ideas central to relevant adaptation works which address these limitations and have motivated our work.

1.4.1 Sources of domain shifts in medical imaging

In UDA, there is an underlying source domain distribution $\mathfrak{p}_s(x, y) \in \mathfrak{p}_S$ and a different target domain distribution $\mathfrak{p}_t(x, y) \in \mathfrak{p}_T$. A labeled dataset $\mathcal{D}_S$ is then sampled i.i.d. from $\mathfrak{p}_s(x, y)$, and an unlabeled dataset $\mathcal{D}_T$ is sampled i.i.d. from the marginal distribution $\mathfrak{p}_t(x)$.

Table 1.1 A summary of possible domain shifts, given a source distribution $\mathfrak{p}_s(x, y)$ and different target distribution $\mathfrak{p}_t(x, y)$

Covariate shift $\mathfrak{p}_s(x) \neq \mathfrak{p}_t(x)$	Label shift $\mathfrak{p}_s(y) \neq \mathfrak{p}_t(y)$
Conditional shift $\mathfrak{p}_s(x \mid y) \neq \mathfrak{p}_t(x \mid y)$	Concept shift $\mathfrak{p}_s(y \mid x) \neq \mathfrak{p}_t(y \mid x)$

Four different forms of domain shifts may be distinguished, as seen in Table 1.1. Existing works largely concentrate on a single shift, sometimes making the explicit assumption that all other shifts are null across domains.

1.4.1.1 Covariate shifts

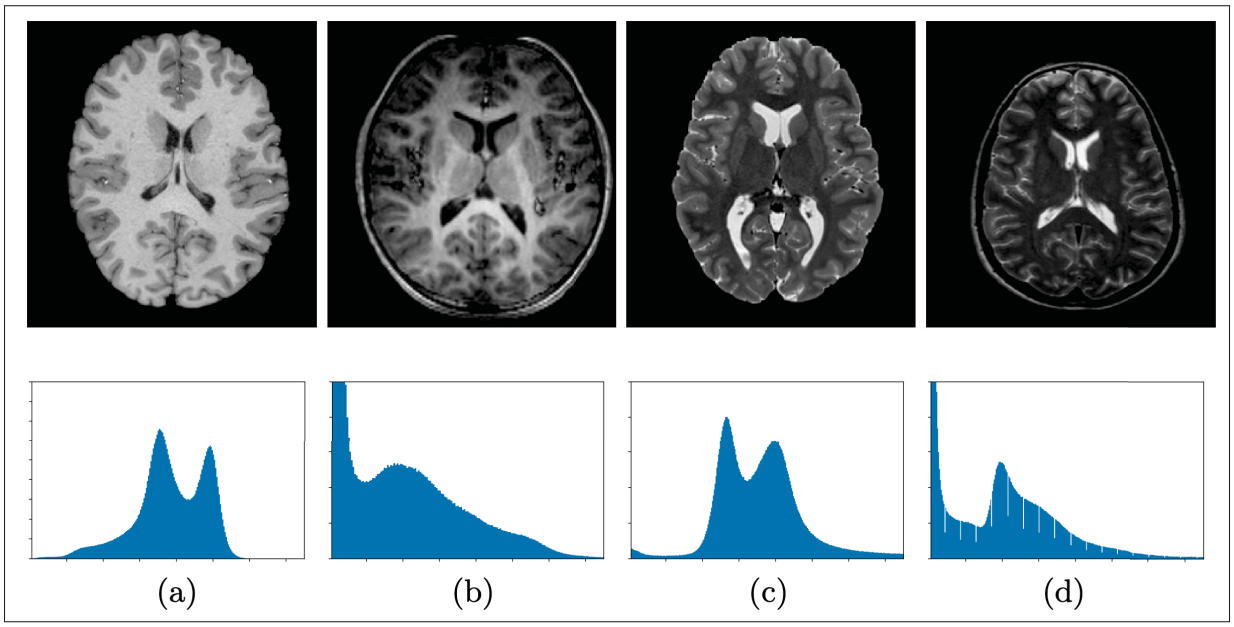


Figure 1.8 Illustrations of the covariate shift. Image slices (top) and corresponding intensity distribution (bottom) of normalized T1-weighted (a, b) and T2-weighted (c, d) MRIs from different scanners. Image source: Karani *et al.* (2021)

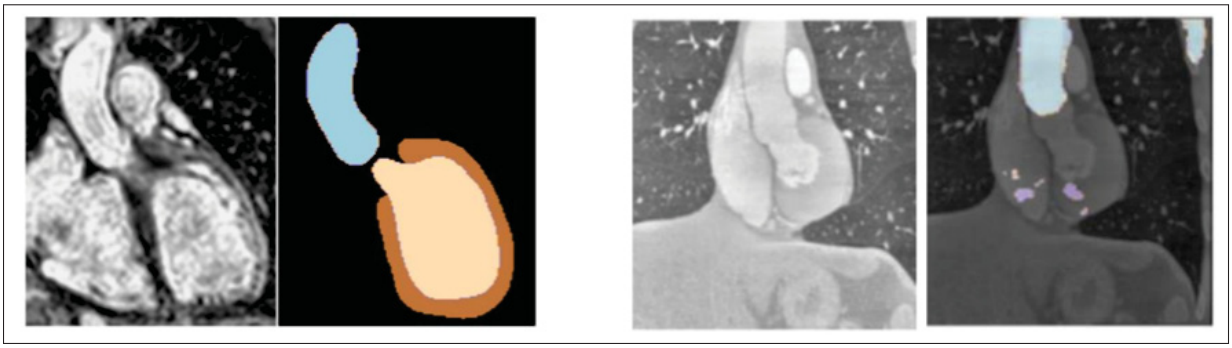


Figure 1.9 Illustration of the sensibility of CNN to cross-modality domain shifts: a CNN segmentation network trained on cardiac MRI images is able to produce excellent segmentation predictions on in-domain samples (left), while it fails to segment the cardiac structures on CT samples (right)

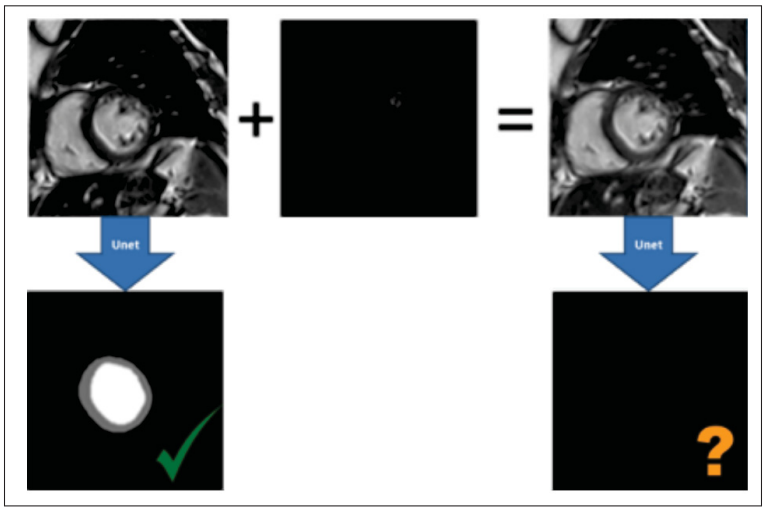


Figure 1.10 Illustration of the sensibility of CNN to small perturbations : a CNN segmentation network trained on cardiac MRI images completely fails to detect the cardiac structures on the modified image, while the perturbation is imperceptible to the human eye. Image source: Yan *et al.* (2019)

1.4.1.1.1 Acquisition-related shifts

Covariate shifts are the most commonly studied domain shifts. They represent a misalignment of the source and target samples' marginal distribution. In medical imaging, acquisition-related

differences are common sources of covariate shifts. They can be divided into two categories. First, in single modality DA, the source and target domains share the same data modality. In this case, the distribution shift can stem from: different vendors (Philips, GE, Siemens), machines (drift in SNR over time, gradient non-linearities), or even change in protocols on the same scanner (flip angle, echo or repetition time, etc.) between the source and the target domain. Second, cross-modality DA is usually more challenging, as it tries to bridge the distribution gap between different imaging modalities (e.g., MRI, CT and PET). For example, the source domain consists of MRI images, whereas the target domain contains CT images.

Despite high-level information similarity, there are considerable differences (e.g. intensity, contrast, noise level) due to these image acquisition shifts. Figure 1.8 shows an example of 2D slices from two T1-weighted and two T2-weighted MRI datasets from different scanners, exhibiting such intensity variations. Note that these variations involve primarily low-level features, e.g., brightness, contrast, texture, resolution, lighting, and color. These shifts often severely degrade the accuracy of CNN, as can be seen in Figure 1.9. Finally, CNN are also well-known to be sensitive to carefully crafted perturbations, i.e. adversarial attacks Yan *et al.* (2019), as illustrated in Figure 1.10.

These acquisition-related shifts will be the main focus of this dissertation.

1.4.1.1.2 Population shifts

Another important source of covariate shifts can occur when a cohort does not adequately represent the range of possible patients and symptoms. For instance, when the source and target datasets have been acquired disjointly or in different sites, substantial shifts in the patient characteristics between the source (or train) and the target (or test) dataset are to be expected. In general, when the demographics of the source population do not match that of the target population, the trained model will present lower performance in the underrepresented groups Sangalli *et al.* (2021). Specifically, distribution shifts in patient age, height, weight, race/ethnicity, gender, clinical diagnosis, severity of underlying conditions for instance will

likely cause performance drops. In fact, these drops reveal unwanted biases that have been introduced into the model, induced by the selection bias in the source dataset. For instance, data imbalance in the training dataset has been found to cause lower performance of cardiac MRI segmentation models for Black patients in Puyol-Antón, Ruijsink, Piechnik, Neubauer, Petersen, Razavi et al. (2021). This may result in downstream biases if diagnostic analyses are performed on the automatically delineated images. Amongst difficulties, medical imaging publications often do not report the demographics of the data. Indeed, while the implications of fairness for the broad field of ML have received an increased attention, surveying and addressing the potential unequal behavior of deep segmentation models in the medical computation community is still a nascent field Lara *et al.* (2022).

1.4.1.2 Conditional shifts, label shifts and concept shifts

Since various classes could have their unique shift protocols, the conditional shift can be used to align the shift of $\mathfrak{p}(x \mid y)$. For instance, as seen in Figure 1.11, while most tissues can look quite similar using different types of CT scans (IV contrast, oral contrast, non contrast), important abnormalities and pathologies will only become apparent when using contrast CT. Estimating $p_t(x \mid y)$ without $p_t(y)$ is, however, an ill-posed problem Zhang, Schölkopf, Muandet & Wang (2013). Another source of discrepancy is the label shift, where the sample fraction of classes labels in $K = \{1, \ldots, K\}$ varies across the two domains. Finally, the concept or annotation shift can arise when annotations of medical images exhibit bias from the annotators, for e.g. due to their competence levels. This annotation shift is, however, a less critical issue in most medical imaging tasks. Note that in realistic clinical settings, a combination of the four shifts described would be involved.

1.4.1.3 Link between causality and domain shifts

As we have seen in Figure 1.10, even without any domain shift clearly perceptible to the human eye, a deep model can fail on target images. In fact this can reveal the model's inability to distinguish between correlations and causation, and its use of spurious correlations in its

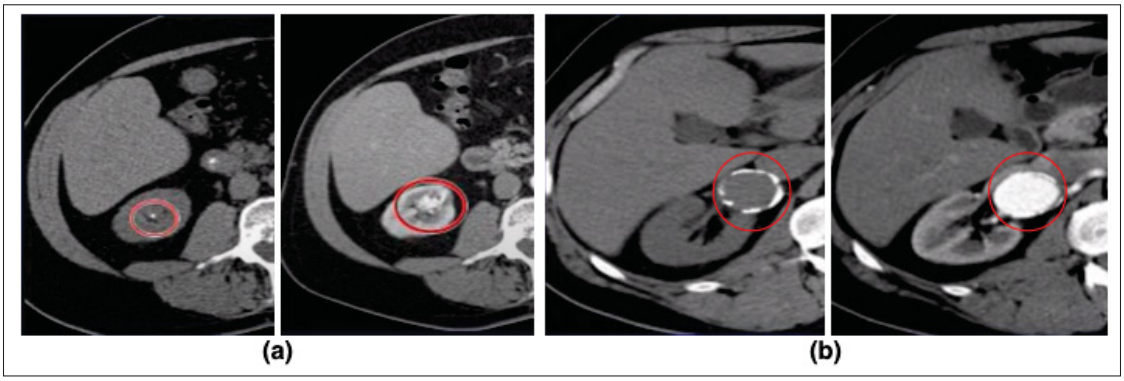


Figure 1.11 Illustrations of the conditional shifts with non-contrast versus contrast CT. (a) arteriovenous malformation is invisible in the non-contrast CT (left) versus apparent with contrast (right). (b) aneurysm is invisible in the non-contrast CT (left) versus apparent with contrast (right)

predictions Vlontzos *et al.* (2022). A telling example is that of the many methods claiming to be able to diagnose COVID-19 using chest X-rays, but eventually falling short since they were detecting fictitious correlations, such as hospital IDs or the ethnicity of the patient, as revealed in DeGrave, Janizek & Lee (2021). Instead, much like humans experts, a model that has learned a correct causality representation should not be sensitive to domain shifts. As such, the main hypothesis of causal inference in medical imaging is that causality aware methods can learn to account for the shifts and biases such as those described above, and reduce their effects Vlontzos *et al.* (2022). Conversely, forcing models to be robust to domain shifts can eliminate spurious correlations and instead help them focus on true causal relationships. For instance, Chen, Wei, Kumar & Ma (2020c) analyzed a setting where spurious features correlate with the label in the source domain but are independent of the label in the target. They showed that self-training methods such as in Section 1.3.3.2 are able to remove such spurious correlations in the adaptation model.

The next section introduces the domain adaptation problem.

1.4.2 Domain Adaptation : problem statement and notations

Source Training Phase

We consider a set $\mathcal{S}$ of source images $I_s : \Omega_s \subset \mathbb{R}^d \rightarrow \mathbb{R}$, $d \in \{2, 3\}$, $s = 1, \ldots, S$. The ground-truth K-class segmentation of I_s is available, and can be written, for each pixel (or voxel) $i \in \Omega_s$, as a simplex vector $\mathbf{y}_s(i) = (y_s^1(i), \ldots, y_s^K(i)) \in \{0, 1\}^K$. Additionally, we consider a set $\mathcal{T}$ of images in the target domain, $I_t : \Omega_t \subset \mathbb{R}^2 \rightarrow \mathbb{R}$, $t = 1, \ldots, T$. A performance metric P is chosen to quantify the quality of segmentation models.

We aim to transfer knowledge learned from $\mathcal{S}$ to $\mathcal{T}$ to perform the semantic segmentation task on $\mathcal{T}$. The first step involves training a deep segmentation network $\mathcal{N}(\cdot; \theta)$ on the source domain only, by minimizing a standard supervised loss ℓ with respect to network parameters θ:

$$\mathcal{L}_s(\theta, \Omega_s) = \frac{1}{S} \sum_{s=1}^{S} \frac{1}{|\Omega_s|} \sum_{i \in \Omega_S} \ell\left(\mathbf{y}_s(i), \mathbf{p}_s(i, \theta)\right) \tag{1.53}$$

where $\mathbf{p}_s(i, \theta) = (p_s^1(i, \theta), \ldots, p_s^K(i, \theta)) \in [0, 1]^K$ is the softmax output of the network at i in image I_s.

At the end of the source training phase, the network function is $\mathcal{N}(\cdot; \tilde{\theta})$ with $\tilde{\theta}$ tuned to perform well on the source domain.

Adaptation Phase

The adaptation phase is then initialized with the network parameters $\tilde{\theta}$ obtained from the source training phase. The goal of domain adaptation is to design a framework able to produce an adapted network function $\mathcal{N}(\cdot; \tilde{\theta}_2)$ with $\tilde{\theta}_2$ the parameters tuned to perform on the target domain. Although no clear definition exists of a successfully adapted network, an intuitive condition is that the performance on the source dataset and target dataset are similar:

$$P(\mathcal{N}(\cdot;\tilde{\theta}),\mathcal{S}) \approx P(\mathcal{N}(\cdot;\tilde{\theta}_2),\mathcal{T}) \tag{1.54}$$

This can be done by minimizing the supervised loss ℓ with respect to network parameters θ on the target domain:

$$\mathcal{L}_t(\theta,\Omega_t) = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{|\Omega_t|}\sum_{i\in\Omega_T}\ell\left(\mathbf{y}_t(i),\mathbf{p}_t(i,\theta)\right) \tag{1.55}$$

This setting, referred to as supervised fine-tuning, needs target labels to retrain the model with the supervised loss. However, the target domain is often unlabeled or only weakly labeled, so the ground-truth segmentation labels $\mathbf{y}_t$ of the target images I_t are inaccessible. Therefore, the supervised loss in the target domain $\mathcal{L}_t(\theta,\Omega_t)$ cannot be computed.

As a result, the common strategy for adapting segmentation networks is to introduce an optimization problem of the following general form:

$$\min_{\theta}\mathcal{L}_s(\theta,\Omega_s) + \lambda\mathcal{L}_{st}(\theta,\Omega_s,\Omega_t) \tag{1.56}$$

where $\mathcal{L}_s(\theta,\Omega_s)$ is the main task loss on the source domain from the source training phase, $\mathcal{L}_{st}(\theta,\Omega_s,\Omega_t)$ is the adaptation or the cross-domain loss, and λ is a weighting parameter. The adaptation loss $\mathcal{L}_{st}(\theta,\Omega_s,\Omega_t)$ can take many different forms, which can be regrouped into two categories. 1) domain-divergence minimization: the loss can be a measure of the divergence of the model's behavior on the source and the target domain, or a domain translator of the input images; 2) regularization methods, penalizing improbable segmentation predictions in the target domain, or encouraging specific properties of the segmentation predictions. *The methods proposed in this dissertation fall in this second category.*

Two remarks on this DA problem statement:

- The underlying assumption for keeping the first objective $\mathcal{L}_s(\theta, \Omega_s)$ in the adaptation phase is that the network should keep performing well on the segmentation task on the source domain, as this will help it perform the same task on the target domain. However this is not always guaranteed, as exploiting the source domain undesirably can degrade the performance in the target domain : this is called negative transfer Zhang, Deng, Zhang & Wu (2022).

- The DA problem statement does not state specific conditions with regard to the final performance in the source domain $P(\mathcal{N}(\cdot; \tilde{\theta}_2), \mathcal{S})$. In fact, DA methods do not often report this performance, and some works have shown the catastrophic forgetting of adapted models, i.e. low $P(\mathcal{N}(\cdot; \tilde{\theta}_2), \mathcal{S})$ Volpi, de Jorge, Larlus & Csurka (2022).

We present hereafter two simple and common strategies to improve the robustness of models to domain shift. If by themselves, they still remain inadequate to successfully adapt segmentation networks in many applications, they are often used in combination with more complex methods.

1.4.3 Simple Domain Adaptation techniques

Data Augmentation

Amongst the simplest strategy to improve a model's robustness to domain shift, data augmentation techniques can be used to diminish the gap between source and target domain performance. Specifically, the domain shift can be simulated by conducting transformations on the source or the target data. For instance, Orbes-Arteaga, Varsavsky, Sudre, Eaton-Rosen, Haddow, Sørensen et al. (2019) encourages prediction consistency between the original and the augmented version of images in the target domain. Physics-driven data augmentation is proposed in Jog, Hoopes, Greve, Van Leemput & Fischl (2019) to enforce invariance to the data generation process. Alternatively, Volpi, Namkoong, Sener, Duchi, Murino & Savarese (2018) generates worst-case transformations under the current model to improve its performance on such adversarial examples.

Adapting Batch Normalization (BN)

In Vanilla BN, the features from the source and target domains are pooled into a single batch and fed as the input of the BN layer. In this way, BN shares its sufficient statistics across the two domains, which is clearly sub-optimal due to the domain shift. To address this issue, Li *et al.* (2016) separates the BN source and target statistics during training. Hu, Uzunbas, Chen, Wang, Shah, Nevatia et al. (2021b); Nado *et al.* (2020) estimate target statistics during testing to improve generalization - contrary to the standard approach for in-distribution test images, where the global statistics of all training samples are used to normalize the test data.

The field of adapting segmentation networks is currently dominated by DA adversarial models. We can broadly group such methods into adversarial discrepancy and adversarial generative methods, which we review in the next section.

1.4.4 Adversarial Adaptation

1.4.4.1 Adversarial discrepancy

Adversarial discrepancy methods generally use an adversarial objective to promote domain uncertainty concerning a domain discriminator. Specifically, two networks, the main task network and domain discriminator network, are trained concurrently. The discriminator tries to distinguish between the source and the target domain, while the task network performs the original task in both domains. It is one of the most common discrepancy minimization techniques for adapting classification networks Shu, Bui, Narui & Ermon (2018); Tzeng, Hoffman, Saenko & Darrell (2017).

Extending these ideas for the segmentation of images from different domains, works such as Dou *et al.* (2019); Hong, Wang, Yang & Yuan (2018); Tzeng *et al.* (2017) propose to alternate the training of two networks, one learning a discriminator between source and target features and the other generating segmentations for both domains. However, the dimensionality of the learned features and the label space in a segmentation task is much higher than in classification

tasks. This can hinder learning a discriminator boundary between the source and target domains in the feature space. To address this issue, Tsai *et al.* Tsai *et al.* (2018) proposed to minimize the discrepancy between the domains in the output space, instead of the feature space. The underlying motivation is that the output space conveys domain-invariant information about segmentation structures, for instance, shape and spatial layout, even when the inputs across domains are substantially different. Finally, a line of works combines matching the features and labels in natural Hoffman, Tzeng, Park, Zhu, Isola, Saenko et al. (2018); Zhang, Qiu, Yao, Liu & Mei (2018a) and medical images Chen, Dou, Chen, Qin & Heng (2020b).

1.4.4.2 Adversarial generative

Another main branch of adversarial methods follows a generative approach e.g., via generative adversarial networks (GANs) Goodfellow et al. (2014). The vanilla GAN is a generative model designed to draw samples from the desired data distribution without the need to explicitly model the underlying probability density function. It consists of two neural networks: the generator G and the discriminator D. The generator is expected to generate samples that have visual similarity with real samples, while the discriminator is expected to indicate the probability of its input being a real or fake sample. Figure 1.12 shows an illustration of a vanilla GAN framework.

In domain adaptation, GANs have been used to transform the images from the source domain so that they look ‘’similar” to those from the target domain, or vice-versa Yi *et al.* (2019). When aligned pairs are not available, image-to-image translations can be trained by combining two generators together, and a cycle-consistency loss is introduced, such as in CycleGAN Huo et al. (2019); Zhu *et al.* (2017); alternatively, some methods Bousmalis, Silberman, Dohan, Erhan & Krishnan (2017); Liu & Tuzel (2016) integrate a task-specific regularization to GANs. These image-level adaptation approaches can be more suitable for segmentation tasks to preserve the original structure information of the pixels/voxels. Moreover, they need less *labeled* data than other methods, as GANs can use semi-supervised learning for training. They have been successful both in computer vision Bousmalis *et al.* (2017); Hoffman *et al.* (2018); Russo, Carlucci, Tommasi & Caputo (2018); Sankaranarayanan, Balaji, Castillo & Chellappa (2018);

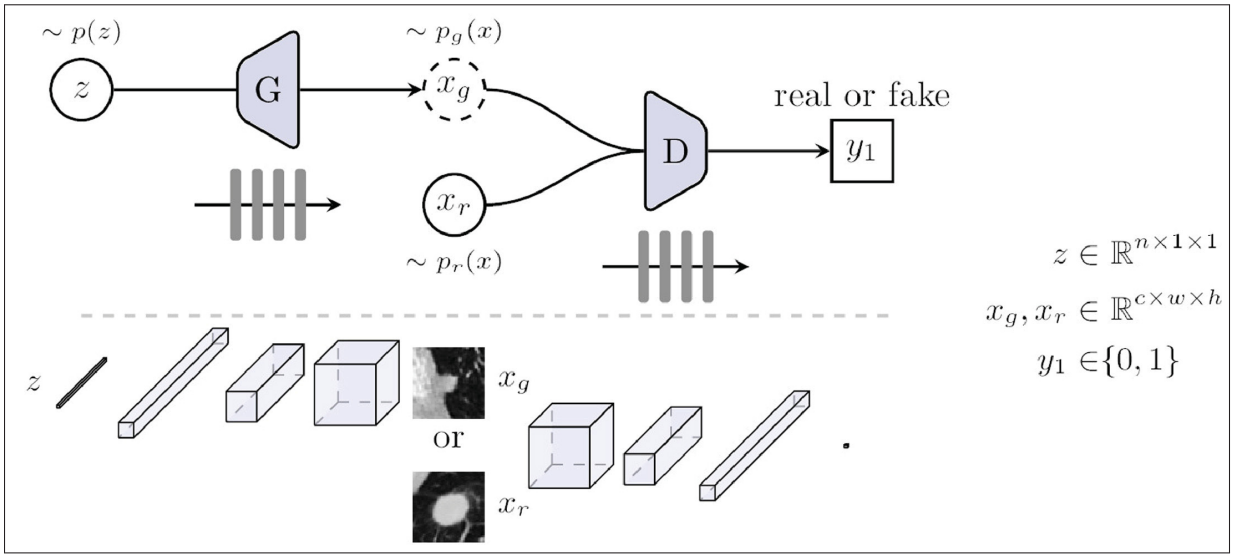


Figure 1.12 Schematic framework of the vanilla GAN for the synthesis of lung nodules on CT images. Top of the figure shows the network architecture. The bottom part shows the input, output and internal feature representations of the generator G and discriminator D. Image source: Yi *et al.* (2019)

Zhu *et al.* (2017) and medical imaging Chen, Dou, Chen & Heng (2018a). Figure 1.13 shows an example of unpaired image-to-image translation via adversarial generative networks proposed in Zhu *et al.* (2017). However, a major drawback is that these methods are optimized in ignorance of the downstream task such as segmentation; important information can be lost or distorted, e.g. anatomy hallucination in CycleGAN Cohen, Luck & Honari (2018); similarly, the destruction or introduction of lesions, is a prohibitive risk which limits the application of CycleGAN in many medical applications.

1.4.4.3 Limitations of adversarial adaptation

Adversarial adaptation has become a prevalent method for adapting segmentation networks, both for natural images Chen, Li & Van Gool (2018c); Hoffman *et al.* (2018); Hong *et al.* (2018); Tsai *et al.* (2018) and for medical imaging Gholami et al. (2018); Javanmardi & Tasdizen (2018); Kamnitsas et al. (2017); Zhao et al. (2019). However, some critical drawbacks should be highlighted, and can prohibit the use of such methods in many realistic clinical applications:

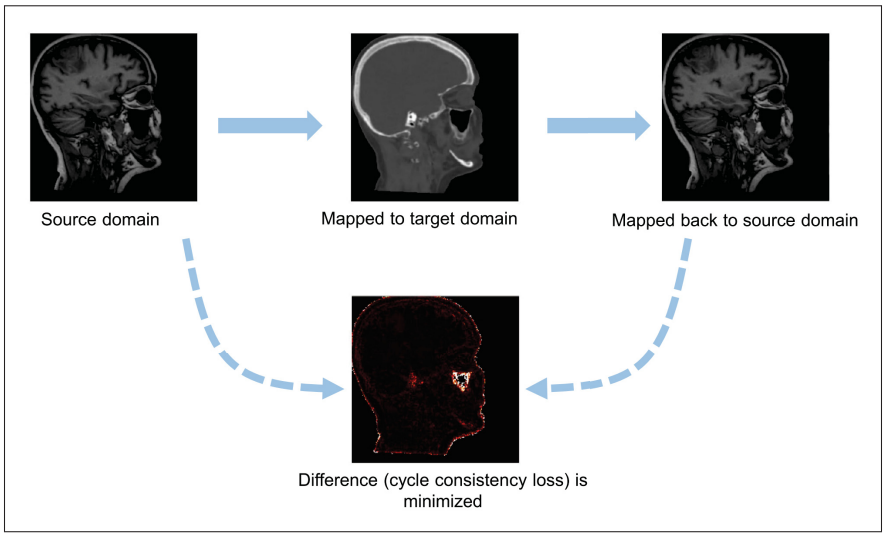


Figure 1.13 Image-to-image translation via cycle consistency loss (CycleGAN) Zhu *et al.* (2017). The source domain MR image is mapped to the target domain CT image, and then mapped back to the source domain. The difference between the input MR image and the reconstructed MR image is minimized. Image source: Guan & Liu (2021)

- **ill-suited to small datasets:** as noted in Zhao, Yue, Zhang, Li, Zhao, Wu et al. (2020), adversarial approaches tend to under-perform on small datasets, because of the dependence of the two competing networks to converge on a min–max game. For adversarial generative methods, large datasets are required to prevent the discriminator from overfitting to the training examples, otherwise, its feedback to the generator becomes useless and the training diverges Yi *et al.* (2019).

- **training and computational complexity:** the optimization procedure for adversarial methods is complex, as it alternates the training of two networks, the discriminator and the generator or task network. Thus it is often unstable and based on many heuristics, making these models challenging and time-consuming to train successfully. Additionally, complex approaches like CycleGAN Zhu *et al.* (2017) must be trained independently for each new test domain, which is computationally expensive and requires a significant amount of source and test data.

- **ill-suited to the segmentation task:** adversarial domain adaptation aims at learning a domain-invariant representation, by learning a discriminator boundary between the source and target domains and aligning target samples to the source domain. However, as pointed out in a few works in computer vision Shu *et al.* (2018); Zhang *et al.* (2020a); Zou *et al.* (2018) one should be cautious with using adversarial training for high-capacity models, as is the case for segmentation. Indeed, Shu *et al.* (2018) verified empirically that it was possible for a deep network to achieve a small source error and a small feature divergence, without performing well on the target task.

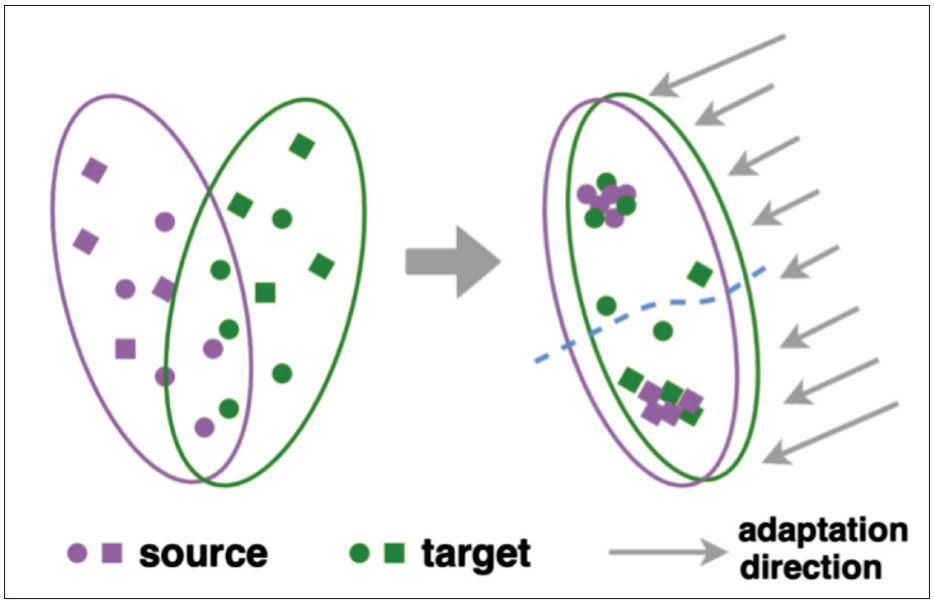


Figure 1.14 Risk of biased domain adaptation. The left and right parts represent the distributions of source and target domains before and after adaptation respectively. The majority of DA techniques focus on matching target samples to well-clustered source classes, which causes misalignment close to the classifier boundary and transfers unnecessary source-specific knowledge to the target. The decision boundaries also remain unclear for target samples close to the boundaries. Image adapted from Hu *et al.* (2022)

- **biased domain adaptation:** Hu *et al.* (2022) highlights another flaw of adversarial DA methods. Specifically, the transferability of a model is often estimated by the target domain score on the source domain classifier, leading to a significant bias in favor of the source

domain. In other terms, under the supervision of source labeled samples, the source domain distribution is well trained and little influenced by the target one, whereas the target domain distribution tends to match the source domain rather than promoting the source to align the target one. As a result, in addition to learning domain invariant knowledge (features), this leads to adapting unneeded source-specific knowledge to the target domain, i.e., biased domain adaptation. See Figure 1.14 for an illustration of the risk of biased DA.

- **unrealistic setting:** finally and most importantly, recall that in many medical applications, the synchronized use of source and target images may be prohibited Kaissis *et al.* (2020). By essence, adversarial methods cannot be used in these settings, as they explicitly minimize the divergence between domains. Thus, computing their cross-domain loss $\mathcal{L}_{st}(\theta, \Omega_s, \Omega_t)$ always requires access to images from both source and target domains. Therefore, even if effective, adversarial methods can be inapplicable when the joint access to source and target images is denied.

Given the limitations of domain adversarial methods identified above, we present below alternative constraints that can be placed on segmentation models to achieve domain adaptation.

1.4.5 Self-supervision for adapting segmentation networks

1.4.5.1 Self-training and Entropy minimization

Extending from self-training methods in semi- and weakly- supervised learning (see Section 1.3.3.2), self-training frameworks have been proposed for adapting segmentation networks in a new unlabelled target domain. For instance, Zou *et al.* (2018) generates pseudo-labels from the most confident predictions in the target domain. Yet this can lead the adaptation model to be biased towards the initially well-transferred classes and to disregard other difficult classes. To prevent this issue, Zou *et al.* (2018) introduces class-balanced self-training, with class-wise confidence levels.

Closely-related to the self-training procedure, entropy minimization is a common loss used in the adaptation of segmentation networks. It starts from a simple observation: models trained only on the source domain tend to produce over-confident, i.e., low-entropy, predictions on source images and under-confident, i.e., high-entropy, predictions on target ones. This is illustrated in Figure 1.15. It is observed that prediction entropy maps of anatomical structures from the source domain look like edge detection results with high entropy activations only along the borders of the structure. On the other hand, predictions on target images are very uncertain, resulting in noisy outputs with high entropy activation maps. This is the motivation for enforcing high prediction certainty (low-entropy) on target predictions as well. The Shannon entropy minimization takes the following form:

$$\mathcal{L}_{\text{ent}}(p_{\boldsymbol{\theta}}) = -\sum_{k}\sum_{i\in\Omega} p^{k}(i,\boldsymbol{\theta})\log(p^{k}(i,\boldsymbol{\theta})) \tag{1.57}$$

The entropy loss above can be seen as a soft-assignment version of the pseudo-label cross-entropy loss in Section 1.3.3.2. Therefore, it also has a strong link to the supervised task, making it a principled loss for unsupervised domain adaptation. Different to self-training, entropy minimization does not require complex scheduling procedures. One of the advantages of entropy is that it is general across tasks, and has been successfully used for adapting classification networks Wang *et al.* (2021) as well as segmentation ones Vu, Jain, Bucher, Cord & Pérez (2019); Wu, Zhang, Zhou, Yang, Zhao & Latecki (2020). Interestingly, contrary to a common self-training assumption, Vu *et al.* (2019) stated that training on the "hard" or "most confused" pixels can produce better performance.

However, entropy minimization exhibits the same drawbacks as self-training, namely instability of the optimization and risk of collapse to a trivial solution. To address this problem, a common solution is to further add a class balancing prior and/or a spatial prior in the target domain, such as is done in Vu *et al.* (2019). Alternatively, Wang *et al.* (2021) stabilizes entropy minimization by adapting only a small subset of network parameters, while the rest of the network is frozen, as we detail in Section 1.4.5.4.

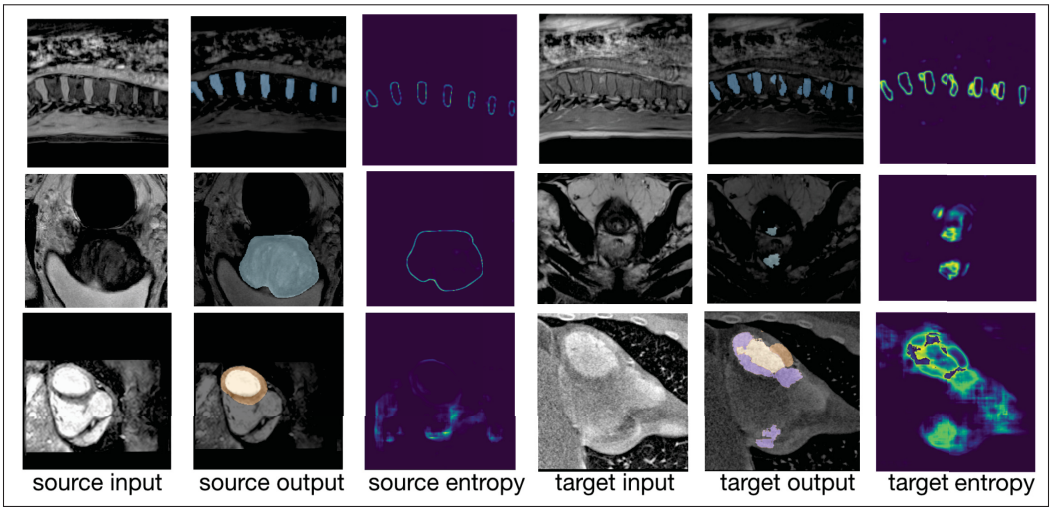


Figure 1.15 Visualization of severe domain shifts between source and target modalities along with their corresponding predicted segmentation and entropy maps in three applications. Top: 2 spine images from Water (left) and In-Phase (right) MRI. Middle: 2 prostate MRI images from different sites. Bottom: 2 cardiac images from MRI (left) and CT (right). The domain shift in the target causes a drop in confidence and accuracy

1.4.5.2 Self-supervision with pretext tasks

Both self-training with pseudo-labels and entropy minimization are self-supervised learning tasks directly related to the supervised task in Eq. (1.55) that would be addressed in the ideal scenario with available target-domain annotations. An alternative strategy is to introduce simple auxiliary tasks, i.e. proxy or pretext tasks which are not directly related to the main task, but are hypothesized to indirectly help it.

Proxy tasks derive a self-supervised label $y_t^{'}$ from the input I_t without the task label y_t. The supervision consists in modifying the target data according to known transforms, training the pretext network to predict such transforms; thus, the transforms are the labels for the pretext task, which is solved via an auxiliary network. It is then concatenated with the task-specific network. The former acts as a generic feature extractor, and the latter leverages such features to solve the main task of interest. Sometimes, both networks are fine-tuned, and sometimes the

auxiliary network is frozen while the task-specific one is fine-tuned. Figure 1.16 presents an example of self-supervised adaptation framework integrating a pretext task.

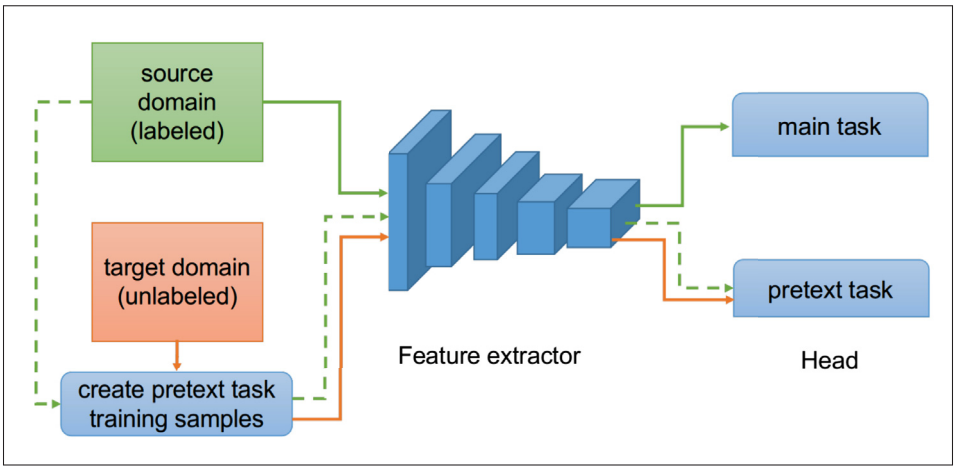


Figure 1.16 Self-supervised domain adaptation with a pretext learning task which can automatically create labels from target domain images. The pretext and main task (semantic segmentation) are learned jointly via multi-task learning. Image source: Xu *et al.* (2019)

A popular proxy task is image reconstruction Bousmalis, Trigeorgis, Silberman, Krishnan & Erhan (2016); Ghifary, Kleijn, Zhang & Balduzzi (2015); Ghifary, Kleijn, Zhang, Balduzzi & Li (2016): in Ghifary *et al.* (2015), for instance, an autoencoder converts images into analogs in several related domains; among other proxies, Bucci, Loghmani & Tommasi (2020) used rotation prediction, Doersch, Gupta & Efros (2015) used patch location prediction, and Xu *et al.* (2019) used both; Carlucci, D'Innocente, Bucci, Caputo & Tommasi (2019) proposed to solve a jigsaw puzzle to address domain adaptation; Pathak, Krahenbuhl, Donahue, Darrell & Efros (2016) reconstructed part of the image with image completion.

A challenge of these methods is that there is no standardized way to select useful pretext tasks for domain adaptation. Too much progress on a proxy task could interfere with performance on the main task, or alternatively the proxy task can be too trivial to induce improvement on the main task, as demonstrated in Kolesnikov, Zhai & Beyer (2019). As such, care is needed in

choosing a proxy task compatible with the domain shift, and to balance optimization between the main task and the proxy task.

1.4.5.3 Curriculum domain adaptation for segmentation

Similar in spirit to self-supervision with proxy tasks, curriculum adaptation for segmentation stems from the observation that some tasks are “easier” and, more importantly, may suffer less from the domain discrepancy than the segmentation task. For instance, instead of trying to directly adapt a segmentation model to a new target domain, Zhang *et al.* (2020a) started by estimating the label distributions $\tau_e(t, \cdot)$ for each image I_t, which they hypothesize is an easier task. This additional task is not self-supervised, but instead trained on the source images, and then inferred on the target images. Different methods can be envisioned to obtain an estimation of the label distribution. In Zhang *et al.* (2020a), image features are extracted from the output of a deep network, and used as the input to shallow models to predict the label distribution.

Once the "easy" task is solved, the "hard" task is tackled, i.e. adapting the semantic segmentation network. To this end, Zhang *et al.* (2020a) used the predicted target-domain label distribution $\tau_e(t, \cdot)$ as additional information to guide the adaptation. Specifically, the segmentation predictions over the target images are constrained to match $\tau_e(t, \cdot)$. The resulting loss for adapting the segmentation network is the following:

$$\mathcal{L}_s\left(\theta, \Omega_s\right) + \frac{\lambda}{T} \sum_{t=1}^{T} KL(\tau_e(t, \cdot), \widehat{\tau}_t(t, \theta, \cdot)) \tag{1.58}$$

where $\mathcal{L}_s\left(\theta, \Omega_s\right)$ is the standard cross-entropy loss on the source domain, and $\widehat{\tau}_t(t, \theta, \cdot))$ is the class-ratio of the segmentation network output prediction for the given image I_t.

The motivation of such a method is to reduce the search space for the adaptation of the neural network parameters by adding a powerful constraint on the target output predictions. As with pretext tasks, a drawback of curriculum domain adaptation is the lack of a methodology for determining which are the "easy" tasks that will be more resilient to domain mismatch. In

fact, comparing the performance of a regression task and a segmentation task is an ill-posed problem. In some applications, if the domain discrepancy is too big, the "easy" task could fail. In the example above, a failure of the easy task would lead to large errors in the estimated target-domain label distribution $\tau_e(t, \cdot)$. As a result, mistakes could reinforce themselves by adapting the segmentation network with wrong information. In Chapter 2, we will see how to alleviate this limitation by incorporating weak supervision.

Updating the whole set of model parameters θ is the common strategy when adapting deep networks. However, there is a risk of over-adaptation, where deep models diverge from their training, even more so adapting networks with missing data. Instead, an emerging idea that we present below consists in freezing part of the parameters, while updating the rest of the parameters.

1.4.5.4 Which network parameters should be adapted and/or shared between the source and the target domain?

In most existing UDA methods, the source and target domains share the whole set of segmentation network parameters $\boldsymbol{\theta}$. Sharing some parameters between the two domains is desirable to capture their common characteristics, and because the target data is unlabeled. However, better adaptation performance could be achieved by separating model parameters which capture domain-specific information from domain-invariant one. Batch normalization (BN) parameters are a natural choice to model domain-specific information. First, as we have seen in Section1.4.3, in Vanilla BN, the mean and variance statistics are shared across the two domains, which is clearly sub-optimal due to the domain shift. In Li, Wang, Shi, Hou & Liu (2018), the BN statistics are simply re-estimated on target samples only at test-time. Going one step further, Chang *et al.* (2019b) aims to capture domain-specific information by introducing domain-specific branches for the BN layers. Each branch estimates BN statistics and learns affine parameters for one of the domains; the other model parameters are shared (see Figure 1.17).

Dou *et al.* (2019) introduces even more domain-specificity, designing an independent encoder for each domain and align their feature distributions in the latent space (see Figure 1.18). In

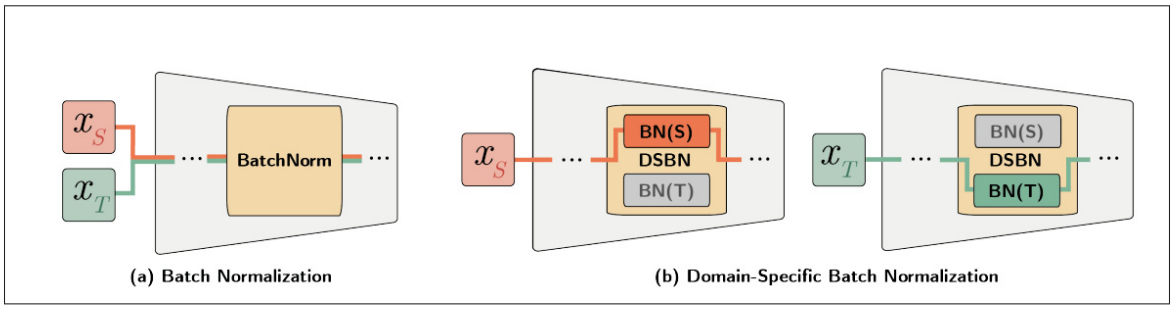


Figure 1.17 Domain-Specific Batch Normalization, which trains two separate domain-specific branches for the BN layers. Image source : Chang *et al.* (2019b)

presence of testing source data, a domain router chooses to use original source early layers; when in presence of testing target data, the router chooses to use target layers. Note that a big disadvantage of this method is the requirement to know which distribution the test data comes from, and its inflexibility to the introduction of other target domains, as they would require other domain-specific modules.

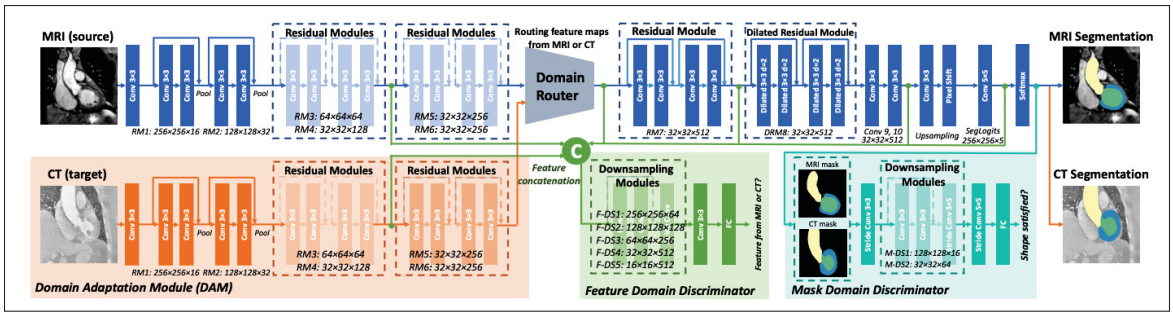


Figure 1.18 The Plug-and-Play AdaNet, with two domain independent encoders (left), a shared decoder (top right) and two discriminators (bottom right). Image source : Dou *et al.* (2019)

In UDA methods described above, some parameters are *shared* between source and target, while others are *domain-specific*. In Rozantsev, Salzmann & Fua (2018), an alternative method is explored, where a two-stream architecture explicitly models the discrepancy between domains, with the weights in corresponding layers being related (via loss terms) but not shared.

1.4.5.4.1 Adapting parameters in scenarios with missing source data

Choosing which parameters to update is even more crucial in UDA with missing source data, such as test-time adaptation (TTA) scenarios which we detail in the next section 1.4.6. Indeed, in these settings, θ is the only representation of the source data, and labels are absent in the target. Over-adaptation could cause models to fail drastically, therefore updating the whole set of network parameters $\boldsymbol{\theta}$ can be inappropriate. In these TTA scenarios with no source data, some parameters are *updated*, while others are *fixed / frozen* to their value at the end of the source training phase.

The TTA method in Wang *et al.* (2021) adapts a model to new target images through entropy minimization in the target domain, and by adapting the BN layers only. Figure 1.19 shows the two steps of their modulations. In their experiments, the combination of the normalization and transformation steps is superior to any single one of these steps, superior to updating the last layer of the model only, and superior to updating the whole set of model parameters θ, which never improves over the unadapted source model. Importantly, their choice of parameters to update also prevents the collapse of entropy minimization collapse to trivial solutions. However, the potential of the method is only demonstrated for small domain shifts.

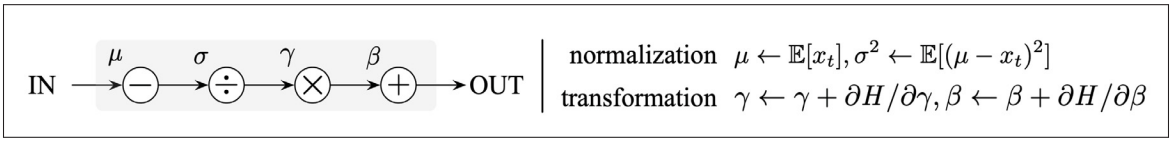


Figure 1.19 Tent Wang *et al.* (2021) modulates batch normalization features by estimating normalization statistics μ, σ, and optimizing transformation parameters γ, β. Tent relies only on the minimization of the entropy H of predictions during testing to adapt a trained network

Amongst other TTA works for classification, Sun, Wang, Liu, Miller, Efros & Hardt (2020b) adapts the convolutional filters of the model, while Liang, Hu & Feng (2020) adapts all but the last layer(s) of the model; on the contrary, Boudiaf, Mueller, Ben Ayed & Bertinetto (2022) only updates the probability outputs of the model but not its trainable parameters.

From these examples we see that identifying which parameters to share versus make domain-specific (in standard UDA) or update versus freeze (in TTA) is still an open problem that involves finding a subset of parameters both expressive and reliable. Moreover, the choice of this subset may well depend on the choice of model architecture, the loss, and tuning.

After introducing relevant domain adaptation methods which have inspired our work, the next section will be dedicated to presenting ideal properties of a domain adaptation framework to ensure its broad applicability in realistic clinical settings. We will aim for our methods to comply with these conditions.

1.4.6 Desirable properties of a domain adaptation framework for broad applicability

A standard source training phase.

Training the task-specific network (e.g. segmentation) with the labeled source data is the main objective of the source training phase. As we have seen in Section 1.4.5.2, many DA methods additionally rely on a pretext or proxy task, e.g., recognizing rotations of an image Sun *et al.* (2020b), denoising segmentation masks (Karani *et al.* (2021)), learning a shape model from the source dataset Yao, Liu, Zhou, Wang, Shen, Yuille et al. (2022). Often, these methods must alter the source training phase, to optimize this proxy loss on the source before adapting to the target domain. While these methods have proved very efficient in tackling unsupervised target domain adaptation, they are not broadly applicable, as they cannot operate simply from an off-the-shelf model trained on the source for the main task only. Moreover, knowing which proxy tasks will help the DA task is not evident, as useful proxy tasks must be both well-defined and non-trivial in the target domain. Given these limitations, we will strive to develop models which adapt segmentation networks without proxy tasks and without altering the source training phase.

No joint use of source and target images

One of the key drawbacks of most DA methods is that they require the joint use of source and target images, in order to compute the cross-domain loss $\mathcal{L}_{st}(\theta, \Omega_s, \Omega_t)$. However, real-world medical applications motivate DA methods that do not assume joint access to the source and target data. First, as the source and target data often come from different institutions, their joint access may be prohibited for privacy, regulatory or bandwidth reasons. Loss and corruption of data is also a frequent impediment. Additionally, for efficiency reasons, it may be computationally impractical to reprocess the source data during the adaptation phase.

These restrictions have motivated *Source-Free Domain Adaptation* (SFDA) Bateson, Kervadec, Dolz, Lombaert & Ben Ayed (2020); Karani *et al.* (2021) such that the adaptation model is independent of the source data given the model parameters. In this setting, the source data is unavailable during the training of the adaptation phase, neither in the form of source images nor ground-truth masks.

In SFDA, the DA optimization problem from Eq. 1.56 is modified to the following general form:

$$\min_{\theta} \mathcal{L}_t(\theta, \Omega_t) \tag{1.59}$$

where $\mathcal{L}_t(\theta, \Omega_t)$ is a self-supervised loss, used as a surrogate to the supervised loss in Eq. 1.55.

No need for a target training dataset.

The standards DA paradigm assumes access to a set of samples from the target distribution during training. Indeed, evaluating DA methods typically encompasses: (i) performing the chosen adaptation strategy on a dedicated training set *Tr* from the target domain; (ii) choosing the final model by measuring performance metrics on a target validation set *Tv* and (iii) measuring the generalization performance on an unseen test set *Te* in the target domain.

However, recent works in the field of *Test-Time Adaptation* (TTA) Sun *et al.* (2020b); Wang *et al.* (2021) argue that it is more efficient to adapt the model directly to the specific subjects from the test set *Te*. Moreover, access to a representative set of samples from the target distribution may be impossible. Indeed, in clinical applications, only a single subject from a new target domain might be given. Standard DA methods cannot operate in this scenario. On the contrary, TTA turns a single unlabeled test subject into a self-supervised learning problem, on which the model parameters are updated before making a prediction. This also extends naturally to data in an online stream. Therefore, the test-time adaptation setting is the most flexible one for adapting deep networks, as it requires the least data. Table 1.2 summarizes the different adaptation settings presented above.

Table 1.2 Adaptation settings differ by their access to data and therefore their losses during training and testing. TTA is the most convenient setting, only needing a single target test sample

setting	source data	target data	train loss	test loss
fine-tuning	-	x^t, y^t	$L\left(x^t, y^t\right)$	-
domain adaptation	x^s, y^s	x^t	$L\left(x^s, y^s\right) + L\left(x^s, x^t\right)$	-
source-free domain adaptation (SFDA)	-	x^t	$L\left(x^t\right)$	-
test-time adaptation (TTA)	-	x^t	-	$L\left(x^t\right)$

To summarize, the conditions above express that for broad applicability, a domain adaptation method should be *robust to missing source and/or target data*. We aim to develop adaptation methods that possess all of these properties. Our final goal will be to adapt a segmentation model to a single target subject with a single network by performing fully test-time adaptation during inference.

CHAPTER 2

CONSTRAINED DOMAIN ADAPTATION FOR IMAGE SEGMENTATION

Mathilde Bateson[1], Jose Dolz[1], Hoel Kervadec[1], Hervé Lombaert[1], Ismail Ben Ayed[1]

[1]École de Technologie Supérieure, 1100 Notre-Dame West, Montreal, Quebec, Canada H3C 1K3

Paper published in *IEEE Transactions in Medical Imaging* (TMI), volume 40.

Presentation

This chapter presents the article "Constrained Domain Adaptation for Image Segmentation", submitted to the journal IEEE Transactions in Medical Imaging, published in July 2021. The initial results were presented as a poster in MICCAI, Shenzhen 2019. The objective of this article is to develop a domain adaptation method based on a constrained formulation, embedding domain-invariant prior knowledge about the segmentation regions. Our general formulation imposes inequality constraints on the network predictions of the target samples, and address the ensuing constrained optimization problem with differentiable penalties, fully suited for conventional stochastic gradient descent approaches. Our results also show robustness to imprecision in the prior knowledge.

2.1 Introduction

Building accurate automatic image analysis systems is a key problem for many biomedical applications. In recent years, Convolutional Neural Networks (CNN) have made a substantial impact and became the de-facto choice in a breadth of computer vision and medical imaging tasks, including classification, detection, semantic segmentation, Dolz, Desrosiers & Ayed (2017); Litjens *et al.* (2017), achieving state-of-the-art or even human-level performances. Nonetheless, CNN typically require a huge quantity of annotated data to perform well. This is especially the case for semantic segmentation, an important first step of the diagnosis and treatment pipeline. To train CNNs for segmentation, pixel or voxel-level annotations of large datasets are commonly used. In 3D medical images, such voxel-level annotations are very cumbersome to obtain, as they require scarce expert knowledge. This has led to many recent efforts, both in computer vision Dai, He & Sun (2015); Pinheiro & Collobert (2014); Wei, Liang, Chen, Shen, Cheng, Zhao et al. (2016) and medical imaging Jia, Huang, Eric, Chang & Xu (2017); Kervadec *et al.* (2019b); Rajchl, Lee, Oktay, Kamnitsas, Passerat-Palmbach, Bai et al. (2016), to develop methodologies mitigating the lack of full annotations, such as semi- or weakly-supervised models Bai, Oktay, Sinclair, Suzuki, Rajchl, Tarroni et al. (2017); Dai *et al.* (2015); Tang *et al.* (2018a); Zhou, Wang, Tang, Bai, Shen, Fishman et al. (2019b).

Typically, segmentation ground-truth is available for very limited data, and the performances of supervised models might drop significantly with new unlabeled samples (target data) that differ from the labeled training samples (source data). These under-performances are a major drawback of standard deep learning techniques, which impedes their deployment in practical scenarios with domain shifts. In medical imaging, such domain-shift scenarios occur frequently when acquisition machines come from different vendors and clinical sites, or when images are acquired across multiple protocols, such as Computed Tomography (CT) and Magnetic Resonance Imaging (MRI) modalities. Acquiring images across different modalities is often very useful to capture different physical properties, which play complementary roles in the clinical procedures for disease diagnosis and treatment. In Figure 2.1, the lower spine is shown across two distinct MRI modalities (Water and In-Phase), and the heart across MRI and CT. In both

cases, one can observe significant differences in contrast, intensity histograms and demarcation between the structures in the two modalities.

Domain adaptation (DA) techniques aim at learning models robust to such distribution shifts. Early deep-learning approaches for domain adaptation investigated minimizing a distance between the features learned from the source and target domains, e.g., the Maximum Mean Discrepancy (MMD) work in Tzeng, Hoffman, Zhang, Saenko & Darrell (2014), thereby aligning the source and target distributions in the latent feature space. With the recent success of generative adversarial networks (GANs) Goodfellow *et al.* (2014), adversarial learning has become the dominating choice for domain adaptation Bousmalis *et al.* (2017); Chen *et al.* (2018a); Hoffman *et al.* (2018); Russo *et al.* (2018); Sankaranarayanan *et al.* (2018).

Unlike adversarial methods, we introduce our Constrained Domain Adaptation to guide the network learning with domain knowledge, e.g., anatomical or learned priors. By enforcing inequality constraints on the network output in the target domain, our method enables to implicitly match prediction statistics between source and target domains, without the burden of two-step adversarial training such as in GANs. Moreover, based on inequality constraints, our framework allows uncertainty in the domain knowledge and can therefore leverage weak labels of the target samples, for instance, in the form of image-level tags for segmentation tasks.

2.1.1 Related work

Domain adaptation is currently attracting substantial research efforts, both in computer vision Hoffman *et al.* (2018); Tsai *et al.* (2018); Tzeng *et al.* (2017) and medical imaging Cheplygina, de Bruijne & Pluim (2019); Kamnitsas *et al.* (2017); Ren et al. (2018); Zhang et al. (2018b). The first attempts to tackle domain shifts were proposed in the vision community, for classification tasksShu *et al.* (2018); Tzeng *et al.* (2017). When the labels are not available for the target domain, the problem is referred to as unsupervised domain adaptation (UDA), and is often formulated as a domain-divergence minimization. These methods rely on the minimization of a discrepancy between distributions, and can be performed at various levels. For instance,

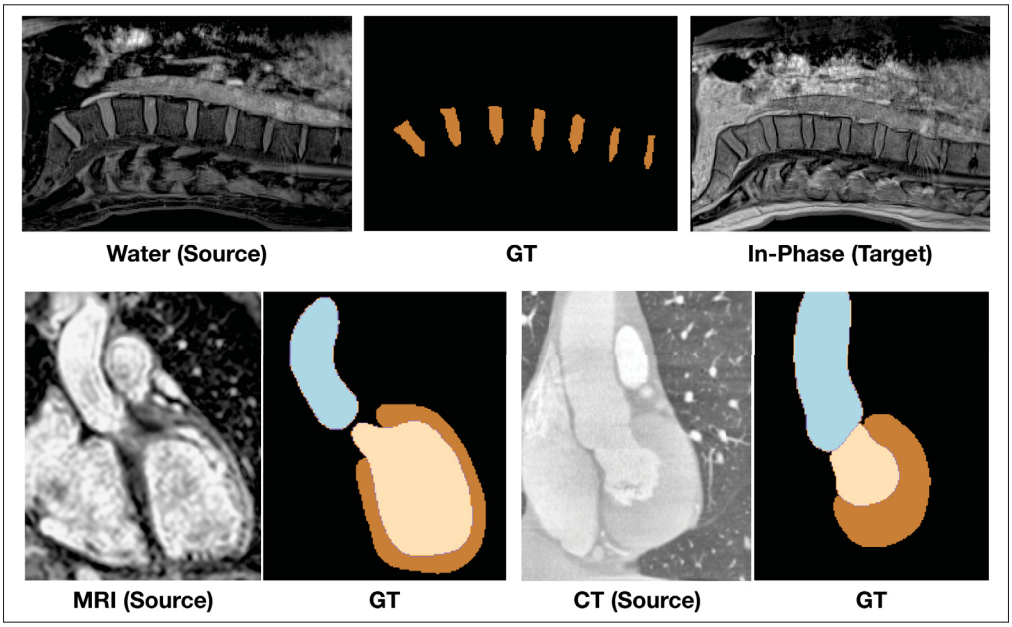


Figure 2.1 Visualization of severe domain shifts between source and target modalities in two applications. Top: 2 aligned spine images from Water and In-Phase MRI and the corresponding ground-truth segmentation, with the intervertebral disks depicted in brown and the background in black. Bottom: 2 cardiac images from MRI and CT, and their ground-truth segmentations. The cardiac structures of AA, LVC and and MYO are depicted in blue, purple and brown, respectively

they could be deployed in the input space, transforming the images from the source domain so that they look ''similar" to those from the target domain, or vice-versa. This approach has enabled a new line of works in both computer vision Bousmalis *et al.* (2017); Hoffman *et al.* (2018); Russo *et al.* (2018); Sankaranarayanan *et al.* (2018) and medical imaging Chen *et al.* (2018a), but still remains inadequate when the domain shift is too large, which may limit its broad applicability. Such a discrepancy minimization could also operate on the representation learned by the CNN. This amounts to aligning the intermediate features from both domains Ghifary *et al.* (2016); Kamnitsas *et al.* (2017); Long, Cao, Wang & Jordan (2015b); Long, Zhu, Wang & Jordan (2016), potentially helping the learned representation to be both useful for classification and invariant with respect to domain shifts Chen & Konukoglu (2018). The most common discrepancy minimization technique uses an adversarial formulation: a discriminator

tries to distinguish between the source and the target domain, while the classifier performs the original task of classifying images from both domains.

Beyond image classification tasks, for which excellent performances were reported Shu *et al.* (2018); Tzeng *et al.* (2017), there is a rapidly growing interest in adapting segmentation networks Kamnitsas *et al.* (2017); Opbroek, Ikram, Vernooij & de Bruijne (2014); Tsai *et al.* (2018), as building pixel-level labels for each new domain is even more tedious. A recent body of work extended domain adaptation ideas for the segmentation of images from different domains Tajbakhsh *et al.* (2020). Most of the studies adapting segmentation networks, either for medical Gholami *et al.* (2018); Javanmardi & Tasdizen (2018); Kamnitsas *et al.* (2017); Khalili, Turk, Zreik, Viergever, Benders & Išgum (2019); Zhao *et al.* (2019) or natural images Chen *et al.* (2018c); Hoffman *et al.* (2018); Hong *et al.* (2018); Tsai *et al.* (2018) use adversarial training. The latter alternates the training of two networks, one learning a discriminator between source and target features and the other generating segmentations. However, the dimensionality of the learned features and the label space in a segmentation task is much higher than in classification tasks. This might invalidate the assumption that the source and target share the same representation at all the abstraction levels of a deep network. To address this problem, Tsai *et al.* Tsai *et al.* (2018) proposed to minimize the discrepancy in the softmax-output space, outperforming feature-matching techniques for unsupervised adaptation in the context of color image segmentation. The underlying motivation is that the output space conveys domain-invariant information about segmentation structures, for instance, shape and spatial layout, even when the inputs across domains are substantially different.

Despite the clear benefits of adversarial techniques in DA for classification Tzeng *et al.* (2017), our experiments suggest that adversarial training may not be well suited to adapt segmentation networks. As pointed out in a few recent work in computer vision Zhang *et al.* (2020b); Zou *et al.* (2018), learning a discriminator boundary between the source and target domains is much more complex for segmentation, as it involves predictions in an exponentially large label space. Instead, it has been shown that self-training Zhang *et al.* (2020b), which generates masks of unlabeled target images via the network's own predictions and uses priors on the spatial layout

of the segmentation regions, can yield better performances. In an approach related to our work Zhang *et al.* (2020b), but investigated for color images, the authors showed that a curriculum learning strategy, which minimizes a Kullback–Leibler (KL) divergence between image-level distributions, for instance, region proportions, can be more effective than adversarial techniques. Finally, it is worth mentioning the recent classification study in Shu *et al.* (2018), which argued that adversarial training is not sufficient for high-capacity models, as is the case for segmentation. For deep architectures, the authors of Shu *et al.* (2018) showed experimentally that jointly minimizing source generalization error and feature divergence does not yield high accuracy on the target task.

Our study draws upon several recent weakly- and semi-supervised segmentation work Jia *et al.* (2017); Kervadec, Dolz, Granger & Ben Ayed (2019a); Kervadec *et al.* (2019b); Pathak *et al.* (2015); Tang *et al.* (2018a,1), which imposed regularization or prior-knowledge terms on the predictions of deep networks, leveraging unlabeled or weakly labeled data. For instance, the work in Tang *et al.* (2018a,1) showed that regularization losses, in the form of a dense conditional random field (CRF) or a balanced graph clustering loss, can achieve excellent segmentation results using only a small fraction of labeled pixels, approaching full-supervision performances in the context of colour images. Along this same vein of research, the work in Jia *et al.* (2017); Kervadec *et al.* (2019b) incorporated priors on the sizes of the segmentation regions via additional loss functions. In this weakly or semi-supervised setting, the main assumption is that the unlabeled data is assumed to be drawn from the same distribution as the labeled data (i.e., there are no domain shifts), which makes the task less challenging than unsupervised or weakly supervised domain adaptation.

2.1.2 Contributions

We propose a general Constrained Domain Adaptation (CDA) formulation for semantic segmentation, which embeds domain-invariant prior knowledge about the segmentation regions. In medical imaging, such knowledge may take the form of simple anatomical information, for instance, structure size or shape, which can be either known *a priori* or learned from the source

samples via an auxiliary task. Our general formulation imposes inequality constraints on the network predictions of unlabeled or weakly labeled target samples, thereby matching implicitly the prediction statistics of the target and source domains, with permitted uncertainty of prior knowledge. Furthermore, our inequality constraints enable to leverage weak annotations of the target data, for instance, in the form of simple image-level tags. We address the ensuing constrained optimization problem with differentiable penalties, fully suited for conventional stochastic gradient descent approaches. Unlike current two-step adversarial training methods, our formulation is based on a single segmentation network, which simplifies adaptation by avoiding extra adversarial steps, while improving training quality.

We report comprehensive evaluations and comparisons on two public segmentation challenges: the intervertebral-disc MICCAI IVD 2018 and the cardiac substructure segmentation MMWHS 2017 challenge. For the adaptation of a segmentation network from one modality to another, our proposed inequality-constrained formulation yielded significant improvements over the state-of-the-art adversarial domain adaptation method in Tsai *et al.* (2018). Moreover, CDA is much faster than adversarial techniques, as the constraints can be learned offline, while the adaptation phase only requires the training of a single segmentation network. The benchmark provided by Dou *et al.* (2019) shows that our formulation outperforms all state-of-the art adversarial methods on the cardiac dataset, including the one proposed in Dou *et al.* (2019). We further provide a comprehensive experimental analysis of CDA, which confirms its robustness to imprecision in the prior-knowledge information. First, we showed that prior knowledge at the image level, for instance, region size, can be learned and estimated via an auxiliary network. Second, we showed that region size could also be estimated from statistics from the source domain, approaching textbook anatomical knowledge. While these estimations are uncertain, we obtained very competitive results of our segmentation network constrained with such priors. Indeed, our method outperforms the recent curriculum adaptation method in Zhang *et al.* (2020b), which does not allow uncertainty in the prior knowledge, in both applications. In addition to our contributions, our code is publicly available[5].

[5] https://github.com/mathilde-b/CDA

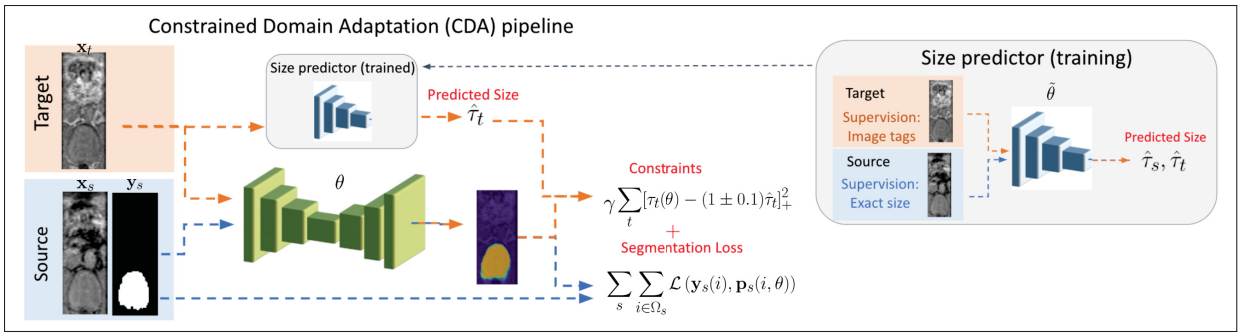


Figure 2.2 (*Left*) Pipeline of the proposed CDA framework. The prior knowledge can be learned and predicted with an auxiliary regression network. (*Right*) The training process of the auxiliary regression network

A preliminary conference version of this work appeared at MICCAI 2019 Bateson, Dolz, Kervadec, Lombaert & Ayed (2019). This journal version provides (1) a broader treatment of the subject with a more detailed description of the method; (2) a new application, the adaptation of cardiac substructure segmentation between MRI and CT images (3) new ablation studies that demonstrate the practical usefulness and robustness of CDA with respect to uncertainty in the prior knowledge, as well as its application to a weak-supervision setting. In particular, we performed comprehensive evaluations for the more realistic and broader setting where prior constraints about the target region are not known/precise, but rather estimated via (a) an auxiliary network, and (b) derived from source statistics, with substantial imprecision.

2.2 Methodology

2.2.1 The proposed Constrained Domain Adaptation

Let us denote $I_s : \Omega_s \subset \mathbb{R}^{2,3} \longrightarrow \mathbb{R}$, $s = 1, \ldots, S$, as the training images of the source domain, each of them having a corresponding ground-truth segmentation, which, for each pixel (or voxel) $i \in \Omega_s$, takes the form of binary simplex vector $\mathbf{y}_s(i) = (y_s^1(i), \ldots, y_s^K(i)) \in \{0,1\}^K$, with K the number of classes, or segmentation regions. If we now consider T images from the target domain, $I_t : \Omega_t \subset \mathbb{R}^{2,3} \longrightarrow \mathbb{R}$, $t = 1, \ldots, T$, we can state domain adaptation for segmentation as

the following constrained optimization problem with respect to the network parameters θ:

$$\begin{aligned} &\min_{\theta} \sum_{s} \sum_{i \in \Omega_s} \mathcal{L}(\mathbf{y}_s(i), \mathbf{p}_s(i, \theta)) \\ &\text{s.t.} \quad f_c(\mathbf{P}_t(\theta)) \leq 0 \quad c = 1, \ldots, C; t = 1, \ldots, T \end{aligned} \tag{2.1}$$

where $\mathbf{p}_x(i, \theta) = (p_x^1(i, \theta), \ldots, p_x^K(i, \theta)) \in [0, 1]^K$ is the softmax output of the network at pixel/voxel i in image $x \in \{t = 1, \ldots, T\} \cup \{s = 1, \ldots, S\}$, and $\mathbf{P}_x(\theta)$ is a $K \times |\Omega_x|$ matrix whose columns are the vectors of network outputs $\mathbf{p}_x(i, \theta), i \in \Omega_x$. In problem (2.1), $\mathcal{L}$ is a standard supervised-learning loss defined solely over the source data (both images and ground-truth masks), e.g., the cross-entropy:

$$\mathcal{L}(\mathbf{y}_s(i), \mathbf{p}_s(i, \theta)) = -\sum_{k} y_s^k(i) \log p_s^k(i, \theta) \tag{2.2}$$

f_c denote the functions that embed constraints on unlabeled or weakly-labeled target-domain images.

Embedding prior knowledge via inequality constraints imposed on the network outputs for target-domain data can be very practical. Assume, for instance, that we have prior knowledge about the size (or cardinality) of the target segmentation region (or class) k. Inequality constraints allow imprecision (or uncertainty) in this knowledge, in the form of lower and upper bounds on region size, unlike Boykov, Isack, Olsson & Ben Ayed (2015); Jia *et al.* (2017); Zhang *et al.* (2020b). For instance, when we have an upper bound a on the size of region k, we can impose the following inequality constraint on the network outputs:

$$\sum_{i \in \Omega_t} p_t^k(i, \theta) - a \leq 0 \tag{2.3}$$

In this case, the corresponding constraint c in the general-form constrained problem in Eq. (2.1) uses the following particular function:

$$f_c(\mathbf{P}_t(\theta)) = \sum_{i \in \Omega_t} p_t^k(i, \theta) - a \tag{2.4}$$

Similarly, we can impose a lower bound b on the size of region k by using the following function instead:

$$f_c(\mathbf{P}_t(\theta)) = b - \sum_{i \in \Omega_t} p_t^k(i, \theta) \tag{2.5}$$

Our framework can be easily extended to more descriptive constraints, e.g., invariant shape moments Klodt & Cremers (2011b), which do not change from one modality to another[6].

An advantage of our formulation is that it can easily integrate weak supervision taking the form of image-level annotations, or tags, in the target domain. Such weak annotations indicate whether a segmentation region k is present or absent in a given image and, therefore, are much less time consuming than full supervision of segmentation, which requires a label for each pixel. Observe that image-level weak supervision can be written conveniently with an inequality constraint in the case of a negative image that does not contain target region k:

$$\sum_{i \in \Omega_t} p_t^k(i, \theta) \leq 0 \tag{2.6}$$

Similarly, for a positive image containing the target region k, we can impose the following constraint:

$$\sum_{i \in \Omega_t} p_t^k(i, \theta) > 0 \tag{2.7}$$

Imposing constraints is common in non-deep image analysis, as it allows to incorporate many types of prior knowledge, such as geometry, context or texture, and has proven effective in many applications, including medical image segmentation Cremers, Osher & Soatto (2006); Klodt & Cremers (2011b); Nosrati & Hamarneh (2016). However, with non-convex deep segmentation models, even when the constraints are convex with respect to the network probability outputs, the general problem in (2.1) is challenging. In standard convex-optimization problems, a common technique to deal with hard inequality constraints relies on the minimization of the corresponding Lagrangian dual, solving primal and dual problems in an alternating scheme

[6] In fact, region size is the 0-order shape moment; one can use higher-order shape moments for richer descriptions of shape.

Bertsekas (1995). For the problem in (2.1), this would involve alternating the optimization of a CNN for the primal with stochastic optimization, e.g., SGD, and projected gradient-ascent iterates for the dual. For semantic segmentation networks involving millions of parameters, this might be computationally intractable. Moreover, the interplay between the primal and dual optimization in the context of deep CNNs might lead to instabilities, seriously affecting the performances of Lagrangian-dual optimization. Therefore, and as pointed out in several recent works Kervadec *et al.* (2019b); Márquez-Neila *et al.* (2017); Pathak *et al.* (2015), despite the clear benefits of imposing hard constraints on CNNs in various applications and problems, such a standard Lagrangian-dual optimization is avoided in the context of modern deep networks.

Instead, in deep networks, equality or inequality constraints are typically handled in a ''soft" manner by augmenting the loss with a *penalty* function He, Liu, Schwing & Peng (2017); Jia *et al.* (2017); Kervadec *et al.* (2019b). The penalty-based approach is a simple alternative to Lagrangian optimization, and is well-known in the general context of constrained optimization; see Bertsekas (1995), Chapter 4. In general, these penalty-based methods approximate a constrained minimization problem with an unconstrained one by adding a term, which increases when the constraints are violated. A disadvantage of the penalty-based approach is that, contrary to Lagrangian optimization, it does not guarantee that the constraints will be satisfied. But it is convenient for deep networks because it removes the requirement for explicit Lagrangian-dual optimization. The inequality constraints are fully integrated with stochastic optimization, as in standard unconstrained losses. This optimization avoids gradient ascent projections over the dual variables, and reduces the computational cost for training. Therefore, in this work, we use a penalty approach, and replace the constrained problem in (2.1) by the following unconstrained one:

$$\min_{\theta} \sum_{s} \sum_{i \in \Omega_s} \mathcal{L}(\mathbf{y}_s(i), \mathbf{p}(i, \theta)) + \gamma \mathcal{F}(\theta) \tag{2.8}$$

where γ is a positive constant and $\mathcal{F}$ a penalty, which takes the following form for the inequality constraints in (2.1):

$$\mathcal{F}(\theta) = \sum_{c=1}^{C} \sum_{t=1}^{T} [f_c(\mathbf{P}_t(\theta))]_+^2 \tag{2.9}$$

with $[x]_+ = \max(0, x)$ denoting the rectifier linear unit function. Clearly, when a constraint is violated, the penalty function is strictly positive; the further we get from the constraint satisfaction boundary, the larger the penalty. A satisfied constraint corresponds to a null cost.

2.2.2 Learning the constraints

An important question is how to best derive useful constraints for CDA. Depending on the application, such priors may be obtained from domain or contextual knowledge, such as anatomical knowledge for medical imaging. For instance, in the first application we tackle in our experiments, we can use human spine measurements that are well known in the clinical literature Berry, Moran, Berg & Steffee (1987) for constraining the sizes of the intervertebral discs in axial MRI slices. When priors are invariant across domains, as is the case for the size of a segmentation region, the statistics of the priors in the source domain can be used. Another option is to train an auxiliary network to learn an estimation of the prior constraints. In our applications below, an auxiliary regression network is trained on the images I_s from the source domain S, where the ground truth size τ_s is known. Then, the learned model is used to predict region-size constraints on the target images. Of course, this might lead to errors in size-constraint predictions over the target images due to the domain shift between the source images used for learning and the target images for inference. To help the regression network, one can add the images I_t from the target domain, using the following "fake" sizes labels for each structure k:

$$\tau_t = \begin{cases} \bar{\tau}_S & \text{if region k is within image } t. \\ 0 & \text{otherwise.} \end{cases} \tag{2.10}$$

where $\bar{\tau}_S$ is the median of ground truth sizes for structure k in the source images I_s. This corresponds to a weakly-supervised setting, where the image-level tag information in the target domain is available. In our experiments, the auxiliary regression network R with parameters $\tilde{\boldsymbol{\theta}}$ is trained to predict the segmentation-region size τ_x, within an image x, with the following squared $\mathcal{L}_2$ loss:

$$\min_{\tilde{\theta}} \sum_{x \in S \cup T} \left(R(x|\tilde{\theta}) - \tau_x \right)^2 \tag{2.11}$$

As we will see in our experiments, our inequality-constraint formulation is robust to imprecision in the prior-knowledge information. While the size prior learned and predicted via an auxiliary network is noisy (uncertain), we obtained very competitive results of our segmentation network constrained with such a prior. We recapitulate the proposed constrained domain adaptation pipeline in Figure 2.2.

2.3 Experiments and results

2.3.1 Experiments set-up

2.3.1.1 Dataset

2.3.1.1.1 IVDM3Seg

The proposed CDA method is first evaluated on the dataset from the MICCAI 2018 IVDM3Seg Challenge[7], a study investigating intervertebral discs (IVD) degeneration. This dataset contains 16 3D multi-modal magnetic resonance (MR) scans of the lower spine, with their corresponding manual segmentations. The 8 subjects were scanned with a 1.5-Tesla MRI Scanner from Siemens using Dixon protocol, at two different stages. In our experiments, models are trained on fully annotated volumes from the Water modality (source domain S), and validated on the In-Phase modality (target domain T). In this setting, the different MRI modalities are acquired from the same patient. To reproduce a more realistic scenario, we have considered that the source and target images are not aligned. This contrasts with the experiments in Bateson *et al.* (2019), where source and target images were registered. From this dataset, 13 scans were used for training, and the remaining 3 scans for validation. As the constraints are imposed axial-wise,

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

we resampled the 36 coronal slices of size 256×256 pixels into 256 slices of 256×36 pixels, as shown in Figure 2.2. Images were normalized between 0 and 1. No additional pre-processing or data augmentation was performed.

2.3.1.1.2 MMWHS

We employed the 2017 Multi-Modality Whole Heart Segmentation (MMWHS) Challenge dataset for cardiac segmentation Zhuang et al. (2019). The data consist of 20 MRI and 20 CT volumes of non-overlapping subjects, with ground-truth masks being provided for both modalities. We aim to adapt the segmentation network for parsing four cardiac structures: the ascending aorta (AA), the left atrium blood cavity (LA), the left ventricle blood cavity (LV) and the myocardium of the left ventricle (MYO). We employed the pre-processed data provided by Dou *et al.* (2019), as well as their data split, with 16 subjects used for training and validation and 4 for testing. In order to obtain a similar field of view for all volumes, they cropped the original scans to center the structures to segment. For each modality, a 3D bounding box with a fixed coronal plane size of 256×256 centered at the heart was used to crop each volume. In Dou *et al.* (2019), a data augmentation based on affine transformations was employed on both the source and target domains for the benchmark (*NoAdap*, *DANN*, *ADDA*, *CycleGAN*, *Pnp-AdaNet*). Similarly, for all the methods we implemented (*NoAdap*, *Adversarial*Tsai *et al.* (2018), *KLAdap*Zhang *et al.* (2020b), *ConsAdap*, *Oracle*), we used a randomized sequence of augmentation steps (contrast shifts, flips) as a data augmentation strategy in the source domain. We did not use any augmentation for the target domain.

Quantitative evaluations and comparisons with state-of-the-art methods are reported. First, to evaluate the impact of domain shift on performance, we compared the proposed loss function to the baselines described in Section 2.3.1.2. We then implemented our proposed CDA in a weakly supervised setting, as described in Section 2.3.1.4. To juxtapose the performance of CDA to other domain adaptation methods under the same conditions, we provide quantitative and qualitative

results of two current state-of-the-art methods for domain adaptation, using an adversarial Tsai *et al.* (2018) or a curriculum strategy Zhang *et al.* (2020b). We also report benchmark results from Dou *et al.* (2019) on MMWHS. Additionally, we provide a comprehensive analysis of the robustness of CDA to imprecision in the prior knowledge. In the first ablation study, we remove the size-regression network, and use source statistics as size priors instead. In the second, we remove the weak image-level tag annotation in the target domain, to test the robustness of our method in a fully unsupervised domain adaptation setting.

2.3.1.2 Baselines

2.3.1.2.1 Lower and upper baselines

To evaluate the impact of the different domain adaptation approaches, we trained a segmentation network in a fully-supervised manner, in both source and target images. Training these fully supervised models reduces to minimizing a standard loss function that evaluates the discrepancy between the CNN predictions and the corresponding ground-truth segmentations: $\min_\theta \sum_{d \in D} \sum_{i \in \Omega_d} \mathcal{L}(\mathbf{y}_d(i), \mathbf{p}_d(i, \theta))$, where D indicates the image domain. Thus, the model trained with the source images, i.e., D=S, will be referred to as *NoAdap*, and will represent the lower baseline. The model employing the target images for training, i.e., D=T, will be denoted as *Oracle*, and will serve as the upper baseline.

2.3.1.2.2 Adversarial domain adaptation

We compared our CDA model to the adversarial approach proposed in Tsai *et al.* (2018), which has demonstrated state-of-the-art performances in the task of unsupervised domain adaptation for natural colour-image scenes. To do so, the penalty $\mathcal{F}$ in Eq. (2.8) is replaced by an adversarial loss, which enforces the alignment between the distributions of source and target segmentations. During training, non-aligned pairs of images from the source and target domain are fed into the segmentation network. Then, a discriminator uses the generated segmentation masks as inputs and attempts to identify the domain of each of these masks (either source or target). In this setting,

we focused on single-level adversarial learning (see Tsai *et al.* (2018) for more details). For a fair comparison, we adopted and improved significantly the performance of the adversarial method in Tsai *et al.* (2018). Following the recent work in Chen, Chen, Chen, Tsai, Wang & Sun (2017b); Pei, Cao, Long & Wang (2018), in the weakly-supervised setting, we boosted the performance of Tsai *et al.* (2018) using exactly the same image-level (tag) class information available to CDA. Specifically, we modified the discriminator loss so as to account for the tag of both source and target images. We experimented with various settings, and found that training the discriminator with only the positive images from both domains increased significantly the performance of Tsai *et al.* (2018). In fact, the use of negative (or mixed) pairs, in which the source and/or target images do not contain the region of interest, confuses adversarial training, reducing its performance in both applications. For the adaptation of the cardiac sub-structure segmentation task, we also report the results by PnP-AdaNet Dou *et al.* (2019), an adversarial method designed specifically for cross-modality DA composed of two independent domain-specific encoders and a decoder, along with a customized network architecture. Dou *et al.* (2019) also provided extensive comparisons to other state-of-the-art adversarial adaptation methods (DANNGanin et al. (2016), ADDA Tzeng *et al.* (2017), CycleGAN Zhu *et al.* (2017)), used in conjunction with the same backbone segmentation architecture as in PnP-AdaNet, which we refer to as AdaNet. These methods are included in Table 2.2 for comparison.

2.3.1.2.3 Kullback-Leibler divergence adaptation

We further compare our approach to the recent curriculum domain method proposed in Zhang *et al.* (2020b), which first learns region-proportion priors, i.e., label distributions. Then, the adaptation phase in Zhang *et al.* (2020b) is based on minimizing a Kullback–Leibler divergence, thereby matching the network predictions' label distributions to these priors. While conceptually related to our approach, the method in Zhang *et al.* (2020b) does not allow imprecision in the priors. Moreover, given the steeper profile of the Kullback–Leibler divergence compared to our penalty function in Eq. (2.9), its optimisation may be less robust than CDA to a noisy prior. For a fair comparison, we used the same prior estimation obtained by the auxiliary regression

network R as in CDA (see Section 2.3.1.4), and adapted the framework in Zhang *et al.* (2020b) to a weakly-supervised setting. Specifically, for each target image, the label distribution prior to be used for Kullback–Leibler divergence matching is:

$$\hat{d}_t = \begin{cases} \frac{1}{|\Omega_t|}\hat{\tau}_t & \text{if region k is within image } t. \\ 0 & \text{otherwise.} \end{cases} \tag{2.12}$$

where $\hat{\tau}_t$ is the predicted size by R on the target image t. For comparability and simplicity, we focused on matching label distributions at the image level, removing the additional superpixel level (see Zhang *et al.* (2020b) for more details). In the experiments below, this method is referred to as *KLAdap*.

2.3.1.3 Supervised Constraints

The proposed CDA method can accommodate both precise and imprecise (or uncertain) prior information about the target region, e.g. size, in the target domain. This is done by imposing inequality constraints of the general form in Eq. (2.1) on the target images. Such inequality constraints could be either tight, when we have precise priors, or loose otherwise. In all the following experiments, we imposed lower and upper bounds for each slice. We trained several models under the same setting, using different constraint values for the size priors on the target images.

First, we investigate the capability of the proposed CDA approach when precise information about the size of the segmentation regions is known. To this end, for each image t and each structure k, of the target domain, we constrained the segmentation size by two prior values, which were derived from the ground-truth size τ_t:

$$\tau_t = \sum_{i \in \Omega_t} y_t^1(i) \tag{2.13}$$

We start by introducing a relatively small uncertainty on this prior information, by adding a ± 10% margin. With the notations from Eqs. (2.4) and (2.5), we have the lower and upper bounds:

$$a, b = \begin{cases} 0.9\tau_t, 1.1\tau_t & \text{if } \tau_t > 0. \\ 0, 0 & \text{otherwise.} \end{cases} \tag{2.14}$$

This setting is later on referred to as $Constraint_{10}$. Then, to evaluate the robustness of the proposed approach to imprecision in the prior knowledge about region size, we also investigated the effect of different levels of tightness of the bounds, by allowing larger margins from the exact size. In particular, we trained additional models with margins on the bounds equal to ± 25%, ± 50%, and ± 75%, which are referred to as $Constraint_{25}$, $Constraint_{50}$ and $Constraint_{75}$, respectively. The aim of this setting is to evaluate how precise the target size information should be. This is different from the main experiments, where the ground truth target size τ_t is unknown.

2.3.1.4 Learning constraints via an auxiliary task

Instead of using bounds derived from the ground-truth size, in this setting, we employ bounds derived from the size estimations produced by the regression model R introduced in Section 2.2.2. In addition to the fully-labeled masks for the source images, we assume that weak image-level annotations are available for the target-domain images. These are image tags indicating whether a given image t contains a region of interest k or not. For each target image, the bounds to be used for adapting the segmentation network with constraints (2.4) and (2.5) are:

$$a, b = \begin{cases} 0.9\hat{\tau}_t, 1.1\hat{\tau}_t & \text{if region k is within image } t. \\ 0, 0 & \text{otherwise.} \end{cases} \tag{2.15}$$

where $\hat{\tau}_t$ is the predicted size by R on the target image t. This setting will be referred to as *ConsAdap* in the following experiments.

2.3.1.5 Deriving constraints from estimated anatomical knowledge

To investigate to possibility to circumvent the auxiliary network, in this setting, we derive the size estimations from the source statistics, which aims at approaching textbook anatomical knowledge. For each 2D target image t and each structure k, the bounds to be used for adapting the segmentation network with constraints (2.4) and (2.5) are:

$$a, b = \begin{cases} 0.9\bar{\tau}_S, 1.1\bar{\tau}_S & \text{if region } k \text{ is within image } t. \\ 0, 0 & \text{otherwise,} \end{cases} \tag{2.16}$$

where $\bar{\tau}_S$ is the median of ground truth sizes for structure k in the source images I_s from the training set. We refer to this ablation study as *ConstraintLit* in the following.

2.3.1.6 Evaluation on Segmentation Performance

In all our experiments, we employed two commonly-used metrics to quantitatively evaluate the segmentation performance of models. First, the Dice similarity coefficient (DSC) which evaluates the degree of overlap between the segmentation regions and the ground truth. Second, the Hausdorff distance (HD) which measures boundary distances. We used the 95-th percentile of the Hausdorff distance (HD95) to mitigate noisy ground truth and/or segmentation regions. Therefore, higher DSC values, and lower HD95 values indicate better segmentation performances. As the data is volumetric, these metrics were computed over the 3D segmentations.

2.3.1.7 Training and Implementation Details

For the segmentation networks, we employed ENet Paszke, Chaurasia, Kim & Culurciello (2016), since it achieves good segmentation performance in a reduced time. We also showed additional results with UNet in order to compare with a different backbone architecture. We employed the standard cross-entropy (CE) for the source segmentation loss, along with results combined with the DiceLoss ($Dice + CE$) Milletari *et al.* (2016); Sudre, Li, Vercauteren, Ourselin & Cardoso (2017). All adaptation models were initialized by training the network with the segmentation

loss only, on the source domain, for 150 epochs. For γ in Eq. (2.8) we adopt a grid search to choose the best value of the weighting γ parameter for each setting. In the adversarial DA setting, we employ the same segmentation network and include the discriminator proposed in Tsai *et al.* (2018), while for *KLAdap*, the Kullback-Leibler divergence in Zhang *et al.* (2020b) was included. In all the domain adaptation experiments, we use Adam optimizer for minimizing the respective loss functions, with an initial learning rate of 1×10^{-3}. The best model was chosen based on the validation set.

Finally, in the *ConsAdap* and *KLAdap* settings, the regression network used to learn the size prior is a ResNeXt 101 Xie, Girshick, Dollar, Tu & He (2017), trained from scratch. We trained it via standard stochastic gradient descent, with a learning rate of 5×10^{-6}. The code is implemented in PyTorch. We ran the experiments on a machine equipped with an AMD Ryzen 1950X 16-Core Processor, 32 GB of RAM and an NVIDIA Titan XP GPU.

2.3.2 Results

2.3.2.1 Quantitative results

Table 2.1 reports the quantitative performance of different methods in spine images. With ENet as the backbone architecture, we observe that *NoAdap* achieves the worst performance, with a 46.8% mean DSC. This is not surprising, since the distributions of source and target images are significantly different due to the presence of the domain shift. This indicates that direct transfer of segmentation models trained on the source cannot handle properly the domain gap. Adopting an adversarial strategy allows to stabilize and improve the results over the lower baseline, achieving a mean DSC of 57.3%. The largest improvement is observed when the domain adaptation strategy incorporates a constrained term on the target predictions. First, we observe in Table 2.3 that if the target size is known, the DSC obtained by $Constraint_{10}$ is 80.4%, which corresponds to 95% of the full supervised model, i.e., *Oracle*. However, knowing with precision the size of

the structure to be segmented is not always feasible. In the more realistic scenario *ConsAdap*, where this size prior is estimated, the mean DSC value only drops to 72.3%, achieving 86% of the performance of the upper bound, *Oracle*. Moreover, our model outperforms *KLAdap* by 3.5%, which uses the same size prior estimation but a Kullback-Leibler divergence as a regularisation loss. This demonstrates the usefulness of using inequality constraints around the estimated prior and a less aggressive loss such as Eq. 2.9. The HD values present a similar pattern across the different models. While the adversarial approach reduced the HD to the half (10.3 mm) compared to the lower baseline model (20.7 mm), *KLAdap* obtained a HD of 6.3 mm. Our proposed model *ConsAdap* further improved the results, achieving a HD of 5.4 mm. To demonstrate that our approach is model-agnostic, and generalizes well to other architectures, we replace ENet by UNet. We observe that the results are consistent with those observed with ENet. In particular, while replacing the segmentation network by UNet brings performance gains across methods, the rankings are maintained, with our approach outperforming prior state-of-the-art models. Last, even though the improvement gap is lesser with UNet, it still shows the effect of domain transfer across the various training settings.

Table 2.2 presents the results for segmentation in cardiac images from the MM-WHS MICCAI Dataset. With no adaptation strategy, the performance of a model learnt on MRI images degrades when it is tested on CT images, with an average DSC of 17.3%. On the other hand, our proposed $ConsAdap$ achieves an average HD of 11.0 mm, and an average DSC of 71.4%, representing 80% of the upper bound model, i.e., $Oracle$, trained on target images. Our method significantly outperforms other state-of-the-art approaches on both metrics. Specifically, the adversarial method in Tsai *et al.* (2018) only yields a 41.1% average DSC and an average HD of 45.9 mm, whereas *KLAdap* obtains closer values, with an average DSC of 70.7%, and an average HD of 12.8 mm. Quantitative results from Dou *et al.* (2019) are also provided, which shows that our method outperforms prior works also on this dataset. It is important to highlight that direct comparison is not appropriate, as the backbone architecture of these methods is different.

Finally, we should note that in both applications, the size estimation obtained by the size regression network R is quite noisy, as shown in Figure 2.3. This suggests robustness to prior imprecision of CDA models as we further explore in the ablation study below.

Table 2.1 Performance comparison of the proposed formulation with different domain adaptation methods for spine segmentation

Backbone	**Methods**	**Target Tags**	**DSC (%) mean ± sd**	**HD95 (mm) mean ± sd**
ENet	NoAdap	×	46.8±11.1	20.7±6.7
	AdversarialTsai *et al.* (2018)	✓	57.3±6.5	10.3±2.6
	KLAdapZhang *et al.* (2020b)	✓	68.8±2.2	6.3±0.5
	ConsAdap (ours)	✓	**72.3±2.6**	**5.4±1.5**
	Oracle	✓	84.5±1.6	3.0±0.3
UNet	NoAdap	×	63.9±7.5	9.6±7.0
	AdversarialTsai *et al.* (2018)	✓	69.0±3.8	6.4±1.5
	KLAdapZhang *et al.* (2020b)	✓	73.3±1.5	5.9±2.0
	ConsAdap (ours)	✓	**73.4±2.4**	**4.7±0.9**
	Oracle	✓	85.4±3.0	2.3±0.3

2.3.2.2 Ablation study on bound precision

We also investigated the impact of prior size imprecision in the target domain on the quality of CDA models. To this end, we increase the lower and upper margins around the true size, as explained in Section 2.3.1.3. Results from this study are reported in Table 2.3 and in Figures 2.6 and 2.7. As expected, having precise size constraints result in higher performing models, close to the full-supervision setting on target images. Nevertheless, allowing large ambiguities on the size of the region of interest (± 25-50%) only degrades the DSC performance by up to 6% on spine images, and to 5% on cardiac images. In the ablation study $ConstraintLit$, we replaced the size regressor by simple source statistics as explained in Section 2.3.1.5. Interestingly, results are well above the baseline for spine and cardiac images, yielding 60.7% DSC and 64.2% average DSC, respectively. This indicates that having a coarse knowledge of the target size can be enough to guide adaptation with CDA. Furthermore, if the target image tag is available, it is possible to circumvent the auxiliary network size regressor R.

Table 2.2 Performance comparison of the proposed formulation with different domain adaptation methods for cardiac segmentation, in terms of DSC (mean±std) and HD (mean±std). (Note: - means that the results are not reported in the original papers)

Backbone	Methods	Target Tags	Myo		LA		LV		AA		Mean	
			DSC	HD	DSC	HD	DSC	HD	DSC	HD	DSC	HD
ENet	NoAdap	×	9.0±11.2	38.6±13.5	26.2±30.2	60.5±16.3	1.5±2.3	36.2±24.0	32.6±37.2	62.8±14.1	17.3±20.2	49.5±17.0
	Adversarial Tsai *et al.* (2018)	✓	19.3±8.3	32.5±10.9	58.1±14.5	54.8±22.4	22.5±18.6	45.9±15.8	64.5±14.8	50.4±8.7	41.1±14.0	45.9±14.5
	KLAdap Zhang *et al.* (2020b)	✓	61.3±6.5	**11.2±5.1**	**77.9±7.3**	21.6±24.1	64.4±10.7	**9.8±2.4**	**79.0±5.6**	**8.5±0.9**	70.7±7.6	12.8±8.1
	ConsAdap(ours)	✓	**61.9±0.9**	16.1±6.5	72.8±12.6	**8.7±2.7**	**73.7±5.7**	10.0±3.1	77.3±6.2	9.0±1.9	**71.4±6.4**	**11.0±3.5**
	Oracle	✓	85.8± 3.3	2.9±1.6	90.1±3.2	4.5±3.0	90.8±3.4	3.0±1.5	91.0±7.3	2.9±1.6	89.4±4.3	3.3±1.9
AdaNet	NoAdap	×	15.3±17.2	-	2.7±0.8	-	3.4±5.8	-	31.5±23.9	-	13.2±11.9	-
	DANN Ganin *et al.* (2016)	×	25.7±13.2	-	45.1±23.6	-	28.3±11.8	-	39.0±35.1	-	34.5±20.9	-
	ADDA Tzeng *et al.* (2017)	×	29.2±16.4	-	60.9±13.2	-	11.2±13.1	-	47.6±15.2	-	37.2±14.5	-
	CycleGAN Zhu *et al.* (2017)	×	28.7±13.3	-	**75.7± 4.3**	-	52.3±21.0	-	73.8±7.4	-	57.6±11.5	-
	PnP-AdaNet Dou *et al.* (2019)	×	**50.8±7.0**	-	68.9± 5.2	-	**61.9±10.7**	-	**74.0±7.3**	-	**63.9± 7.5**	-

Table 2.3 Performance comparison for the proposed formulation with constraints derived from the ground truth ($\text{Constraint}_{25,50,75}$) and from the source-domain statistics (Constraint_{Lit}). ENet is employed as backbone architecture

Dataset	Methods	DSC (%) mean ± sd	HD95 (mm) mean ± sd
IVDM3Seg	NoAdap	46.8±11.1	20.7±6.7
	Constraint_{Lit}	60.7±2.8	7.2±2.4
	Constraint_{75}	65.7±4.2	6.9±1.7
	Constraint_{50}	74.6±2.1	4.5±0.7
	Constraint_{25}	77.5±0.7	4.1±0.5
	Constraint_{10}	**80.4±1.5**	**3.7±0.9**
	Oracle	84.5±1.6	3.5±0.3
MMWHS	NoAdap	38.2±11.4	NA[a]
	Constraint_{Lit}	64.2±6.0	11.6±6.4
	Constraint_{75}	69.4±11.1	10.4±5.9
	Constraint_{50}	79.9±6.8	7.5±2.6
	Constraint_{25}	82.5±6.4	7.5±5.3
	Constraint_{10}	**84.6±5.3**	**7.2±6.0**
	Oracle	89.4±4.3	6.5±5.6

[a]NA means that the value cannot be calculated due to no prediction for that structure.

2.3.2.3 Assessing the impact of the target image tags

We investigated the effect of removing the image-level tag annotation in the target domain. Particularly, we removed the target image tags for both the size regressor and the adaptation phase, as explained in Section 2.3.1.4. Results from this study are reported in Table 2.4. As expected, having image-level tag information considerably helps all the models, which can

Table 2.4 Performance of the different domain adaptation methods obtained when removing the weak image-tag annotations. ENet is employed as backbone architecture

Dataset	**Methods**	**DSC (%) mean±sd**	**HD95 (mm) mean±sd**
IVDM3Seg	AdversarialTsai *et al.* (2018)	48.7±2.4	18.0±7.8
	KLAdapZhang *et al.* (2020b)	52.2±5.9	12.5±4.2
	ConsAdap(ours)	**58.3±2.1**	**7.4±3.1**
MMWHS	AdversarialTsai *et al.* (2018)	38.9±14.6	**32.1±10.4**
	KLAdapZhang *et al.* (2020b)	33.3±13.8	N/A
	ConsAdap(ours)	**49.4±14.8**	42.3±11.5

Table 2.5 Performance comparison of the proposed formulation with different segmentation losses defined over the source spine data. UNet is employed as backbone architecture

Method	**Source Loss**	**DSC (%) mean ± sd**	**HD95 (mm) mean ± sd**
NoAdap	CE	63.9±7.5	9.6±7.0
ConsAdap (ours)	CE	73.4±2.4	4.7±0.9
NoAdap	Dice+CE	65.5±2.1	6.0±0.8
ConsAdap (ours)	Dice+CE	**75.7±1.8**	**4.7±0.7**

be observed from the performance decrease in comparison to the results in Table 2.1 and 2.2. Indeed, the size estimation degrades without the image tag and, as a result, models using a size prior to guide adaptation also see their performance decrease.

An interesting observation in this scenario, however, is the larger gap between the proposed model and prior work, particularly compared to *KLAdap*.

2.3.2.4 Qualitative results

Figure 2.4 and 2.6 depict visual segmentation results for spine images, for the 3 subjects used in validation sets. We visualize the results at the best epoch. It can be seen that without adaptation, the network trained only on source images is unable to recover the 7 distinct IVDs present in all the subjects, and the model trained with adversarial adaptation also struggles (see the second row in Figure 2.4). In contrast, our proposed CDA model - both with supervised and learned constraints - is able to detect the 7 IVD structures in almost all examples. Moreover, the

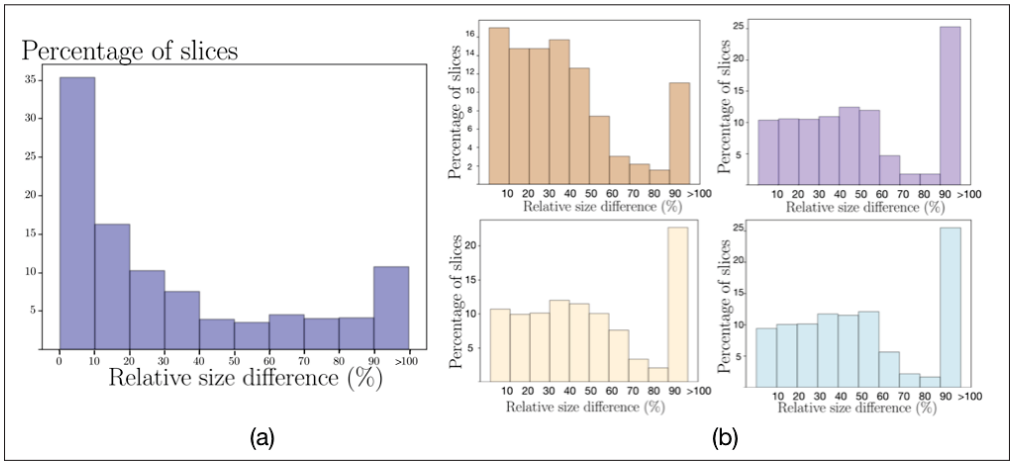


Figure 2.3 Normalized histograms of the relative size difference between ground truth size and size estimated by the auxiliary task in the target domain for spine images (a) and cardiac images (b, clockwise for Myo, LA, AA, LV). This size estimation is used as a prior to guide domain adaptation (see Section 2.3.1.4)

segmentations achieved by all models using the proposed CDA framework have more regular shapes. Figure 2.5 and 2.7 show the visual comparison results on the cardiac dataset, for the 4 subjects in the test set. As illustrated in Figure 2.5, the segmentation results produced by $ConsAdap$ are more similar to the ground truth, in terms of shape and boundary, especially for the MYO and LV structures. Finally, we can visually observe in Figure 2.6 and 2.7 that all constrained models $Constraint_{10}$, ..., $Constraint_{75}$, yield much better segmentations than the lower baseline without any adaptation strategy. Furthermore, as expected, the quality of the segmentations slowly degrades with a more imprecise size prior used for constraining the adaptation.

2.3.2.5 Efficiency

The computational efficiency of constrained formulations, benchmark adaptation formulations and baselines are compared in Table 2.6. The lower (*NoAdap*) and upper (*Oracle*) baselines only need to compute one loss per pass, i.e., cross-entropy, and only use images from the domain on which it is calculated, i.e., source (*NoAdap*) or target (*Oracle*), respectively. As expected,

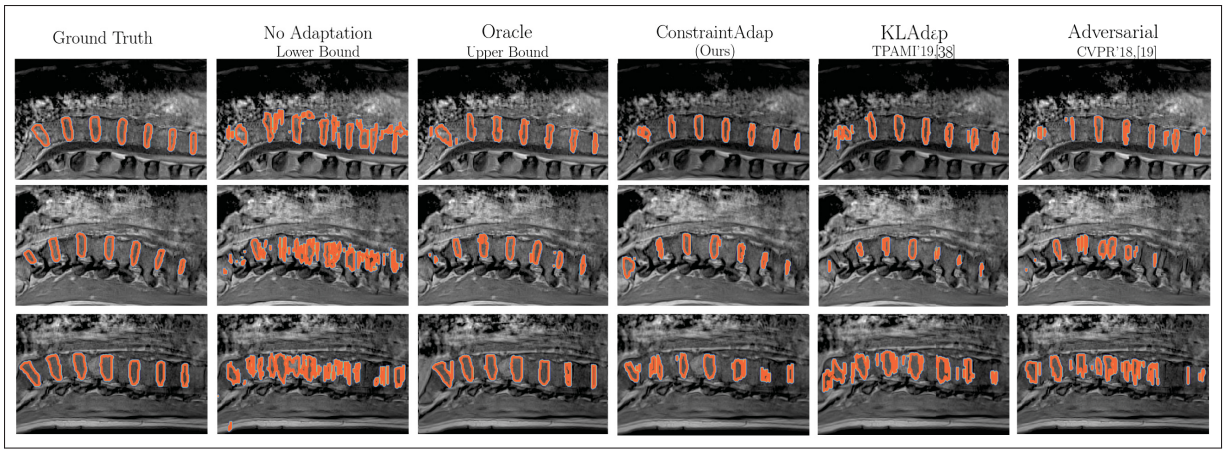


Figure 2.4 Example of the segmentations achieved by our constrained formulation (*ConsAdap*), benchmark models in Zhang *et al.* (2020b) and Tsai *et al.* (2018) and lower (*NoAdap*) and upper baselines (*Oracle*) for intervertebral disks images in the MRI In-Phase modality. Each row shows a different test subject. Images and masks are rotated in the sagittal plane and cropped for better viewing. The IVDs are contoured in red

training times are lower for these methods. All other methods employ images from both domains at each forward pass. Including the quadratic loss with supervised size constraints adds little to the computational time. Using learned priors, such as in models *ConsAdap* and *KLAdap*, does not significantly change the computational time either, even when including the size-regressor training. Particularly, if we consider a two-step process, assuming the same number of epochs for all the models, the proposed constrained framework is still nearly twice faster than the adversarial approach in Tsai *et al.* (2018). The overhead is much higher with the adversarial adaptation, which alternates at each pass between the training of the segmentation network and the training of the discriminator, the latter also requiring inputs from both domains.

Table 2.6 Training times of the various adaptation learning strategies and *Oracle* for a batch size of 12, for spine segmentation

Backbone	**Methods**	**Average Time (s/batch)**
ResNeXt101	*R* (size regressor)	0.2
ENet	NoAdap	0.4
	AdversarialTsai *et al.* (2018)	1.4
	KLAdapZhang *et al.* (2020b)	0.6
	ConsAdap (ours)	0.6
	$\text{Constraint}_{10,25,50,75}$	0.6
	Oracle	0.4

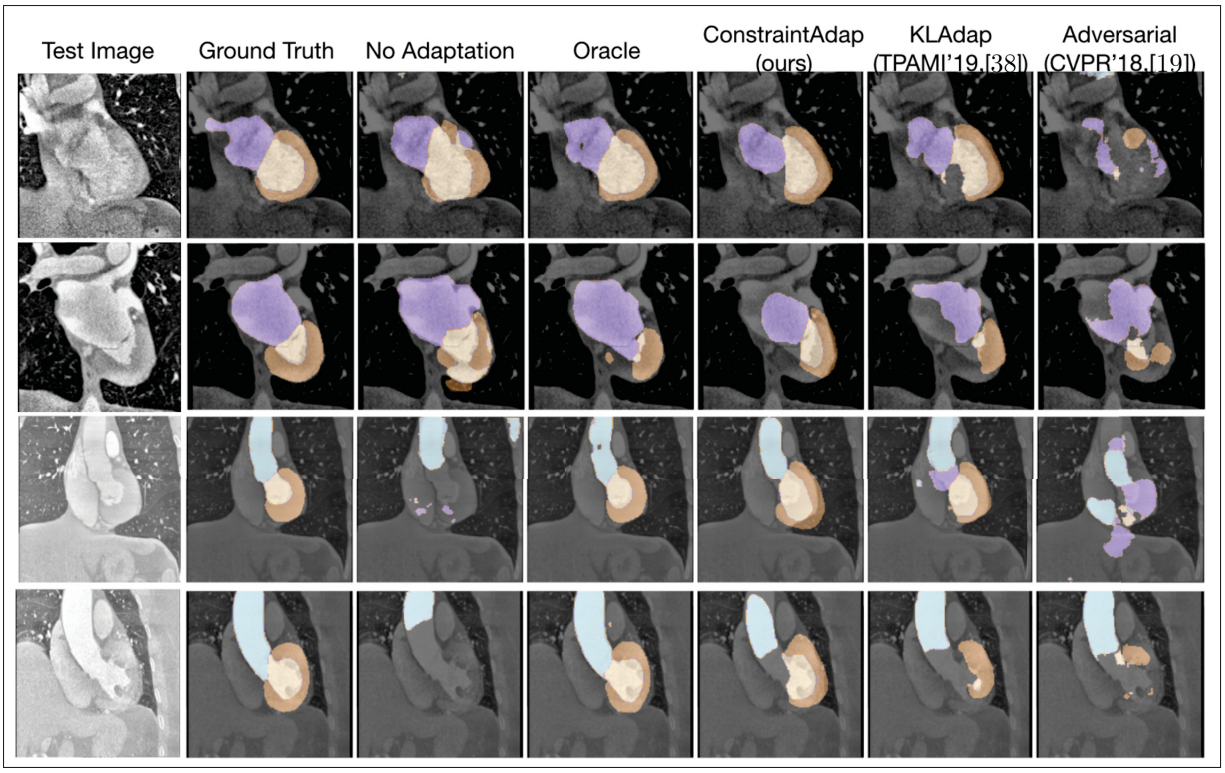


Figure 2.5 Examples of the segmentations achieved by our constrained formulation (*ConsAdap*), benchmark models in Zhang *et al.* (2020b) and Tsai *et al.* (2018) and lower (*NoAdap*) and upper baselines (*Oracle*) for cardiac CT images. The cardiac structures of MYO, LA, LV and AA are depicted in brown, purple, yellow and blue, respectively. Each row shows a different test subject

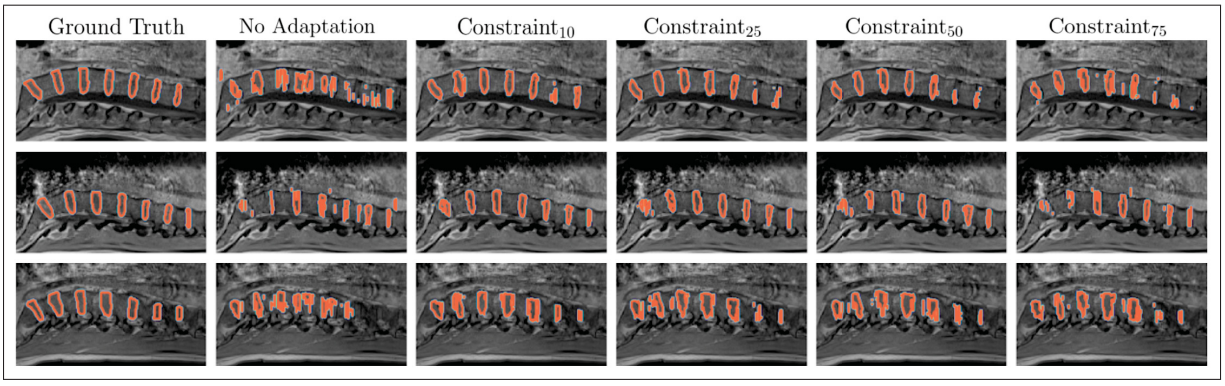


Figure 2.6 Example of the segmentations achieved on spine images by our constrained formulation with tighter to looser constraints (*Constraint*$_{10}$ being the tightest), i.e., increasing prior uncertainty, showing robustness to prior imprecision. Each row shows a different test subject. Images and masks are rotated in the sagittal plane and cropped for better viewing. The IVDs are contoured in red

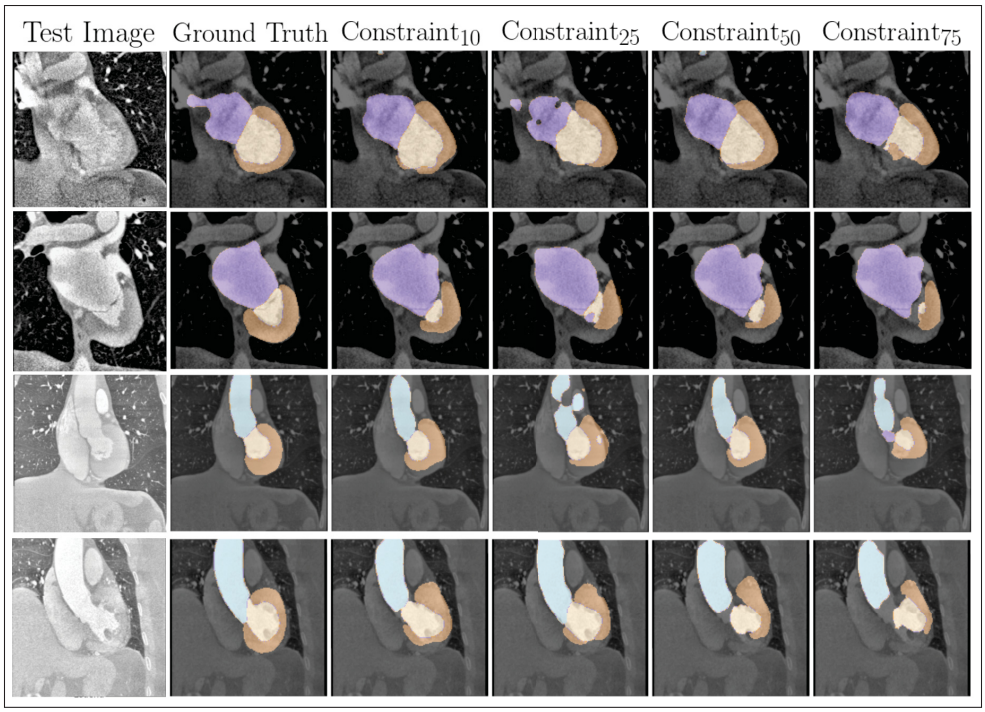


Figure 2.7 Examples of the segmentations achieved on cardiac CT images by our constrained formulation with tighter to looser constraints (*Constraint*$_{10}$ being the tightest), i.e., increasing prior uncertainty, showing robustness to prior imprecision

2.4 Discussion

We presented a method to guide a segmentation network learned on a source domain to perform well on a different target domain, with minimal additional information, for instance, in the form of image-level tags. We showed the versatility of our DA approach, implementing it for drastically different types of images, multi-modal spine MRI images and MRI to CT cardiac images. Our model consistently yields a performance gain of 1-4% in terms of DSC across architectures and datasets, and 4-14% when comparing to state-of-the-art adversarial adaptation approaches. Even though we have evaluated our method on multi-modal (multi-MRI and CT to MRI) spine and cardiac images, it can be applied to other multi-modal scenarios, such as multimodal photoacoustic and optical coherence tomography Hojjatoleslami & Avanaki (2012), for example. Unlike adversarial strategies, which are based on two-step training, our method tackles the adaptation problem with a single constrained loss, simplifying the adaptation of

the segmentation network. In our implementation, the constrained loss matches image-level statistics–the size of the structure to be segmented here–in the target domain through the use of a simple quadratic loss. As demonstrated in our experiments, the performance is significantly improved over the lower baseline. Surprisingly, state-of-the-art adversarial methods Ganin *et al.* (2016); Tsai *et al.* (2018); Tzeng *et al.* (2017); Zhu *et al.* (2017) yield smaller improvements. We hypothesize that this is due, in part, to the difficulty of learning a decision boundary between source and target domains in huge dimensionality. When a very precise size prior is known on the target domain, our framework leveraged this information to improve results up to 95% of the upper bound (the full supervision regime on the target) on two different tasks. As shown quantitatively and qualitatively by our experiments, the structures of interests are much better detected in each patient in the target domain, while the segmentations achieved are greatly improved. Furthermore, we have shown that our method tolerates a substantial imprecision around the true size of structures, and that we can learn a sufficiently accurate size prior with a simple regression network. Although the estimated size prior obtained in our application is quite noisy, and our uncertainty margins very simple, our formulation with learned constraints reaches 86% and 80% of full supervision in spine and cardiac images respectively. We also demonstrate the superiority of our method compared to a domain adaptation model using size statistics matching with a steeper loss and no handling of prior imprecision Zhang *et al.* (2020b).

Arguably, the main limitation of our method relies on obtaining an accurate estimation of region size, which guides segmentation training during the phase of domain adaptation. Learning region size through an auxiliary regression network could be challenging when there is a large shift between the source and target domains. However, we show in Table 2.3 that, even with large ambiguities on size estimation, the performance of the proposed model drops by only 5-6% on both datasets. An interesting finding from our results is that adding the weak image-level class information, i.e., the presence or absence of the target region, for each slice in the target domain greatly helped the auxiliary size regressor network. The need of this weak annotation to approach the performances of full supervision might be seen as another drawback of our method. This contrasts with fully unsupervised domain adaptation methods, which do not require weak

annotations, but are usually unstable and hard to train. Nevertheless, we showed in Table 2.4 that a fully unsupervised version of our method, without access to image-tag information, still outperforms several state-of-the-art adaptation methods based on adversarial training Tsai *et al.* (2018); Zhang *et al.* (2020b). We argue that, despite these drawbacks, our method provides an optimization framework that is simpler and more stable than fully unsupervised adaptation methods.

Future developments could involve a 3D extension, for which some questions related to the incorporation of textbook medical knowledge remain undefined, as it is common to use volumes patches as input to 3D networks Dolz *et al.* (2017). Learning other priors from domain information, such as constraints derived from shape moments Klodt & Cremers (2011b), and better addressing the uncertainty of size estimations, for instance, from deriving more sophisticated margins, are other potential improvements left for future work. For difficult domain adaptation tasks with substantial domain shift, the initial network trained on the source domain may be incapable of detecting any structure in the target domain, complicating the initialization of our method. In such cases, as an alternative approach, weak annotations such as bounding boxes Dai *et al.* (2015); Rajchl *et al.* (2016) in the target domain could be used. Another open question for such difficult applications is the usefulness of enforcing multiple constraints and how to handle the ensuing optimisation problem.

2.5 Conclusion

This study investigated domain adaptation for segmentation with applications for intervertebral discs segmentation in multi-modal MRI and MRI to CT cardiac substructure segmentation. We proposed a constrained formulation for adapting a segmentation network learned on one modality (source domain) to a different modality (target domain), by enforcing image-level statistics in the target domain which we showed could be learned directly from the source domain. Despite its simplicity, the performance of our method comes near that of full supervision with only image-level annotations in the target domain, and very small computation overhead, using basic linear constraints, e.g., target-region size. Extensive experiments demonstrated that our

formulation also outperformed multiple state-of-the-art adaptation methods. Our framework offers, therefore, flexibility, is model-agnostic and opens the door to promising research directions on incorporating a wide variety of new anatomical constraints.

CHAPTER 3

SOURCE-FREE DOMAIN ADAPTATION FOR IMAGE SEGMENTATION

Mathilde Bateson[1], Hoel Kervadec[1], Jose Dolz[1], Hervé Lombaert[1], Ismail Ben Ayed[1]

[1] École de Technologie Supérieure, 1100 Notre-Dame West, Montreal, Quebec, Canada H3C 1K3

Paper published in *Medical Image Analysis* (MedIA), volume 82, September 2022.

Presentation

This chapter presents the article "Source-Free Domain Adaptation for Image Segmentation" submitted to Medical Image Analysis, published in September 2022. An initial article was published in the MICCAI conference, 2020, presented virtually in Lima, Peru. Additionally, the journal article was presented as a short paper poster at the MIDL conference (Medical Imaging with Deep Learning) held in Zurich, Switzerland, 2022. The objective of this article is to introduce a source-free domain adaptation for image segmentation. The unavailability of the source dataset in the adaptation phase is a very frequent DA scenario in medical imaging, where, for instance, the source and target images could come from different clinical sites. To guide adaptation in this context, our formulation is based on minimizing a label-free entropy loss defined over target-domain data. Contrasting with chapter 2, we combined entropy minimization with a domain-invariant prior, in the form of a class-ratio prior on the segmentation regions. We show the effectiveness of our prior-aware entropy minimization in a variety of domain adaptation scenarios, with different modalities and applications, including spine, prostate and cardiac segmentation.

3.1 Introduction

3.1.1 Motivation

Unprecedented advances in visual recognition tasks have been possible thanks to the improvements in hardware, novel deep architectures and availability of large annotated datasets. Deep Convolutional Neural Networks (CNNs) can provide powerful image representations when trained on huge amounts of labeled images, which can be used in a breadth of computer vision problems. For instance, CNNs have outstandingly improved automated methods for segmentation in many natural and medical imaging problems Litjens *et al.* (2017). A major impediment of such supervised models is that they require large amounts of training data built with scarce expert knowledge and labor-intensive, pixel-level annotations. Typically, segmentation ground truth is available for limited data, and supervised models are seriously challenged with new samples (target data) that differ from the labeled training samples (source data). In medical imaging, for instance, the data distribution may vary significantly across different vendors, machines, image modalities and acquisition protocols, as illustrated on Fig. 3.1. Such domain shifts between different scans introduce a significant variability in the appearances of the target regions, impeding the generalization of CNN segmentation models. There has been an ongoing research effort towards improving the performance of models across domains, without retraining them nor labeling entire datasets in new target domains, which would be impractical in medical imaging Cheplygina *et al.* (2019).

Domain Adaptation (DA) addresses the transferability of a model trained on an annotated source domain to another target domain with no, or minimal annotations. With the advent of Generative Adversarial Networks (GANs) Goodfellow *et al.* (2014), adversarial-learning techniques widely dominate the recent literature in domain adaptation for segmentation. One major limitation of adversarial techniques is that, by design, they require concurrent access to both the source and target data during the adaptation phase. More generally, other recent approaches to DA, such as those based on self-training, also use both source and target data during adaptation. However, in many medical imaging scenarios, the source data may not be available in the adaptation phase.

This involves, for example, confidentiality reasons, loss or corruption of the source data, or computational constraints for real-time applications.

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

3.1.2 Related Work

Among the earliest works aiming to address domain-shift problems, Ben-David, Blitzer, Crammer, Kulesza et al. (2010); Crammer, Kearns & Wortman (2008); Pan & Yang (2010) propose to find a mapping of data distributions from a source to a target. More precisely, to tackle the discrepancy between the two domains, the learning process exploits the differences of data distributions across domains, yielding domain-invariant features. The main idea is to find an intermediate feature space where the marginal distribution of the source is similar to the target. Thus, we can assume that, in this intermediate representation, the prediction function is the same across source and target domains. This results in models that can be trained using annotated data sets from the source domain along with unlabeled or weakly labeled target data, with a strong cross-domain generalization ability.

Adversarial methods: Inspired by this assumption, recent works have focused on leveraging deep learning models to extract domain invariant features from input images Ganin & Lempitsky (2015); Long *et al.* (2015b); Tzeng, Hoffman, Darrell & Saenko (2015). Particularly, most of the existing research exploits deep adversarial training Ganin *et al.* (2016) in a wide range of

applications and problems, such as classification Sankaranarayanan *et al.* (2018); Tzeng *et al.* (2017); Van Tulder & de Bruijne (2016); Wachinger et al. (2016) or segmentation Hoffman *et al.* (2018); Huo, Xu, Bao, Assad, Abramson & Landman (2018); Javanmardi & Tasdizen (2018); Kamnitsas *et al.* (2017); Tsai *et al.* (2018); Zhang *et al.* (2018b); Zhao *et al.* (2019). These methods either follow a generative approach, by transforming images from one domain to the other Huo *et al.* (2019); Zhu *et al.* (2017), or minimize the discrepancy in the feature or output spaces learnt by the model Dou *et al.* (2019); Tsai *et al.* (2018); Tzeng *et al.* (2017). As these two perspectives are in essence complementary, the recent methods achieve state-of-the-art performances for adapting semantic segmentation in natural Hoffman *et al.* (2018); Zhang *et al.* (2018a) and medical images Chen *et al.* (2020b) by combining image- and feature-alignment strategies. One major limitation of adversarial techniques is that, by design, they require concurrent access to both the source and target data during the adaptation phase.

Self-training: Amongst alternative approaches to adversarial techniques, self-training Zou *et al.* (2018) and the closely-related entropy minimization Morerio, Cavazza & Murino (2018); Vu *et al.* (2019); Wu *et al.* (2020) were investigated in computer vision. As confirmed by the low entropy prediction maps in Fig. 3.1, a model trained on an imaging modality tends to produce very confident predictions on within-sample examples, whereas uncertainty remains high on unseen modalities. Moreover, the entropy maps can identify inaccurate segmentation regions in these target examples.

As a result, enforcing a higher confidence of predictions in the target domain would help decreasing this performance gap. This is the underlying motivation for entropy minimization, which was first introduced in the contexts of semi-supervised Grandvalet & Bengio (2004) and unsupervised Krause, Perona & Gomes (2010) learning. To prevent the well-known collapse of entropy minimization to a trivial solution with a single class, the recent domain-adaptation methods in Vu *et al.* (2019); Wu *et al.* (2020) further incorporate a criterion encouraging diversity in the prediction distributions, while Bian, Yuan, Wang, Li, Yang, Yu et al. (2020) minimize the uncertainty measured as the variance of the network's output, in combination with adversarial learning. However, similarly to adversarial approaches, all these uncertainty-based methods

require access to the source data, both the images and labels, during the adaptation phase. The source data is used to compute the standard supervised cross-entropy loss and/or used in an adversarial adaptation, to prevent trivial solutions that are obtained by minimizing uncertainty on the unlabeled target images.

Test-time Adaptation: Closest to our work, test-time domain adaptation (TTA) was introduced to improve generalization to new and different data, possibly a single data point, at test times. Most TTA methods comply with the SFDA setting: they relieve the need for accessing source domain data after the source training phase. Initial SFDA attempts addressed adapting classification tasks Liang *et al.* (2020); Nath Kundu, Venkat, Rahul & Venkatesh Babu (2020), either by using generative image translation Benaim & Wolf (2018) or self-supervision Sun *et al.* (2020b); Wang *et al.* (2021). Extensions to segmentation problems He, Carass, Zuo, Dewey & Prince (2020,2); Karani *et al.* (2021) alter the source-domain training with auxiliary branches used to align the target and source domains in the pixel, network-feature, and/or network-output spaces. A drawback of these methods is that the source training phase is non-standard (ex. require training an additional denoising network, Karani *et al.* (2021)) and involve complex training and/or adaptation schemes. Varsavsky, Orbes-Arteaga, Sudre, Graham, Nachev & Cardoso (2020) proposed a test-time adaptation based on domain adversarial learning, which is adapted to a single target-domain subject, but is not source-free.

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

Weakly supervised segmentation in medical imaging: To alleviate the burden of pixel-wise annotation, weakly supervised learning has become a popular strategy. In this setting, the supervision received by the segmentation network may come in the form of image-level tags Ouyang, Xue, Zhan, Zhou, Wang, Zhou et al. (2019); Patel & Dolz (2022); Wu, Du, Luo, Wen, Shen & Feng (2019), bounding boxes Kervadec, Dolz, Wang, Granger & Ben Ayed (2020);

Rajchl *et al.* (2016), points Dorent, Joutard, Shapey, Kujawa, Modat, Ourselin et al. (2021); Khan, Shahin, Villafruela, Shen & Shao (2019), scribbles Tang *et al.* (2018b), target size Jia *et al.* (2017); Kervadec *et al.* (2019b) or, more recently, shape descriptors Kervadec, Bahig, Létourneau-Guillon, Dolz & Ben Ayed (2021). On the one hand, approaches that rely on image-level tags typically use class-activation maps Selvaraju, Cogswell, Das, Vedantam, Parikh & Batra (2017), which are deployed to generate pseudo-labels, mimicking fully-supervised learning. On the other hand, knowledge-driven approaches typically embed prior-knowledge, such as the target size or location, in the learning objective. Furthermore, while most prior literature relies on in-distribution data, a very few attempts investigated domain adaptation in a weakly-supervised setting Bateson *et al.* (2021); Cheplygina *et al.* (2019); Dorent, Joutard, Shapey, Bisdas, Kitchen, Bradford et al. (2020); Paul, Tsai, Schulter, Roy-Chowdhury & Chandraker (2020). These works have shown promising results, especially when dealing with scarce data or severe domain shifts.

Leveraging the target class-ratio as a prior has shown a great potential to guide the training of segmentation models when dealing with limited supervision, including weakly Jia *et al.* (2017); Kervadec *et al.* (2019b), semi-supervised Kervadec *et al.* (2019a); Zhou, Li, Bai, Chen, Han, Wang et al. (2019a) or few-shot Boudiaf, Kervadec, Masud, Piantanida, Ben Ayed & Dolz (2021) learning. In the presence of domain shifts, several recent works have also resorted to this prior as a source of additional supervision Bateson *et al.* (2021); Vu *et al.* (2019); Zhang *et al.* (2020a). An important difference, however, is that prior works require accessing the source data. Indeed, their learning objectives include a cross-entropy loss over the labeled source images during the training of the adaptation phase. This contrasts with our setting, as we relax this requirement.

3.1.3 Contributions

We propose a *Source-Free Domain Adaptation* formulation (SFDA) tailored to a setting where the source data is unavailable, neither its images nor its labeled masks, during the training of the adaptation phase. Instead, our method only requires the parameters of a model previously trained on the source data as an initialization; moreover, it does not use auxiliary branches or additional tasks trained on the source domain, contrary to previous SFDA methods He *et al.*

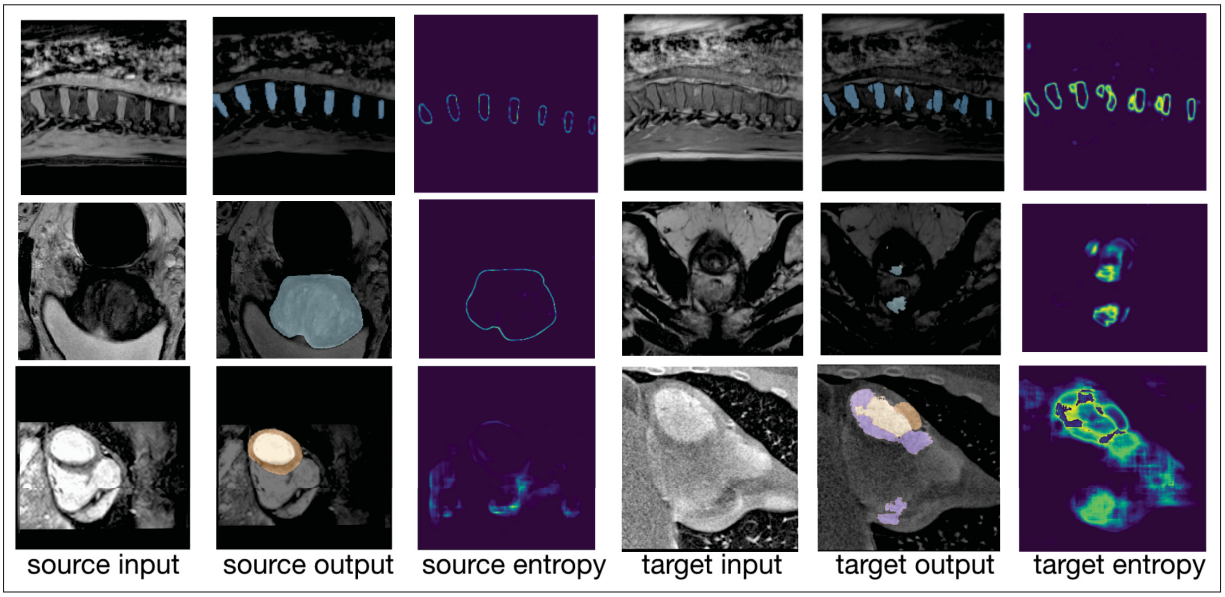


Figure 3.1 Visualization of severe domain shifts between source and target modalities along with their corresponding predicted segmentation and entropy maps in three applications. Top: 2 spine images from Water (left) and In-Phase (right) MRI, with the intervertebral disks depicted in blue and the background in black. Middle: 2 prostate MRI images from different sites. Bottom: 2 cardiac images from MRI (left) and CT (right). The cardiac structures of AA, LV and and MYO are depicted in blue, purple and brown, respectively. The domain shift in the target causes a drop in confidence and accuracy.

(2020,2); Karani *et al.* (2021). Our formulation is based on a minimization of a label-free entropy loss defined over the target-domain data, which we further guide with a domain-invariant prior on the segmentation regions. To facilitate adaptation, we leverage weak supervision in the form of image-level tags in the target domain. Furthermore, we provide an interesting connection between our loss and the mutual information between the target images and their label predictions.

We report a comprehensive set of experiments and comparisons with state-of-the-art domain-adaptation methods, which shows the effectiveness of our prior-aware entropy minimization in three applications: the adaptation of spine segmentation across different MRI modalities, the adaptation of prostate segmentation in MRI modalities across different sites and machines, and the adaptation of cardiac segmentation from MRI to CT. Surprisingly, even though our

method does not have access to the source data during adaptation, it achieves comparable or even better performances than several state-of-the-art methods Dou *et al.* (2019); Ganin *et al.* (2016); Tsai *et al.* (2018); Tzeng *et al.* (2017); Zhang *et al.* (2020a); Zhu *et al.* (2017), while greatly improving the confidence of network predictions.

A preliminary conference version of this work has appeared at MICCAI 2020 Bateson *et al.* (2020). This journal version provides (1) a new loss to tackle source-free adaptation, with an interesting mutual-information perspective and better gradient dynamics than the one introduced in Bateson *et al.* (2020); (2) two new applications; (3) ablation studies; and (4) the introduction of anatomical knowledge to estimate the class-ratio priors, which demonstrates the practical usefulness of our method and its robustness to uncertainty in estimating the priors. Specifically, unlike Bateson *et al.* (2020), we perform comprehensive evaluations in a setting where the class-ratio priors of the target regions are not estimated by an auxiliary network, but rather derived from textbook anatomical knowledge, even with substantial imprecision. We argue that such an approach offers a great potential in multiple clinical settings, particularly when access to source data is compromised. Our framework can be readily used for adapting a breadth of segmentation problems, with the code made publicly available[8].

The contributions of this paper can be summarized as follows:

1. We tackle Source-Free Domain Adaptation (SFDA), a setting where the source data is unavailable, neither its images nor labeled masks, during the training of the adaptation phase. Our formulation allows SFDA with no modification to the source training.

2. We propose a novel loss defined over the unlabeled target-domain data, which integrates the Shannon entropy with a Kullback–Leibler divergence matching the class-ratios of the segmentation regions to an anatomical prior. Furthermore, we motivate our loss with an interesting link to maximizing the mutual information between the target images and their latent labels.

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

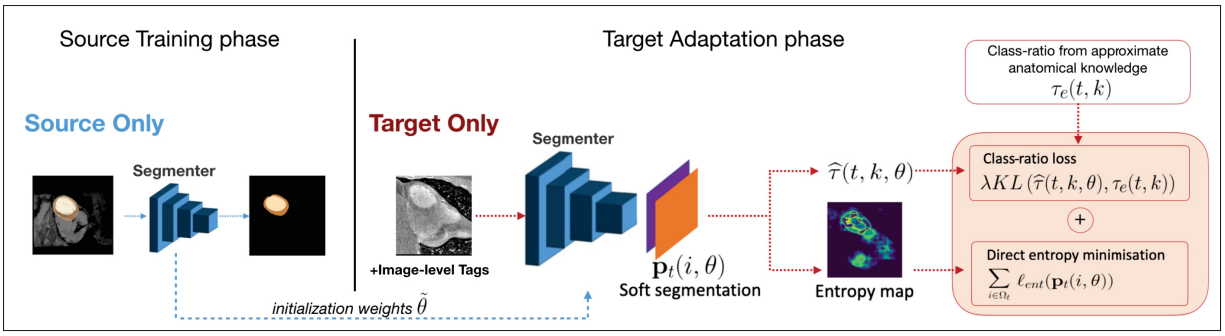


Figure 3.2 Overview of our framework for Source-Free Domain Adaptation: we leverage entropy minimization and a class-ratio prior, to remove the need for a concurrent access to the source and target data.

3. We extensively validate our method on three DA datasets. The results show that our framework can effectively and efficiently address the domain shift problem without accessing the source data during the adaptation phase.

3.2 Method

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

$$\mathcal{L}_s\left(\theta, \Omega_s\right) = \frac{1}{S}\sum_{s=1}^{S}\frac{1}{|\Omega_s|}\sum_{i\in\Omega_S}\ell\left(\mathbf{y}_s(i), \mathbf{p}_s(i,\theta)\right) \tag{3.1}$$

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

The adaptation phase is then initialized with the network parameters $\tilde{\theta}$ obtained from the source training phase. Given a set $\mathcal{T}$ of images in the target domain, $I_t : \Omega_t \subset \mathbb{R}^2 \rightarrow \mathbb{R}$, $t = 1, \ldots, T$,

the first loss term in our adaptation phase encourages high confidence in the softmax predictions of the target, which we denote $\mathbf{p}_t(i,\theta) = (p_t^1(i,\theta), \ldots, p_t^K(i,\theta)) \in [0,1]^K$. This is done by minimizing a weighted Shannon entropy of each of these predictions:

$$\ell_{ent}(\mathbf{p}_t(i,\theta)) = -\sum_k \nu_k p_t^k(i,\theta) \log p_t^k(i,\theta) \tag{3.2}$$

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

However, it is well-known from the semi-supervised and unsupervised learning literature Grandvalet & Bengio (2004); Jabi, Pedersoli, Mitiche & Ayed (2021); Krause *et al.* (2010) that minimizing this entropy loss alone may result into trivial solutions, where the predictions are biased towards a single dominant class. To avoid such degenerate solutions, the recent domain-adaptation work of Vu *et al.* (2019); Wu *et al.* (2020) have integrated a standard supervised cross-entropy loss over the source data, such as in Eq. (3.1), when training during the adaptation phase. This, however, requires access to the source data, both its images and labels, during the adaptation phase. To remove this undesired requirement, we embed a domain-invariant prior knowledge to guide the unsupervised entropy training during the adaptation phase, which takes the form of a class-ratio prior (i.e., the proportion of a region in an entire image). The unknown true class-ratio prior for a class k and image I_t can be computed as follows: $\tau_{GT}(t,k) = \frac{1}{|\Omega_t|}\sum_{i\in\Omega_t} y_t^k(i)$. This gives the size of class k in image I_t over the image size. However, as the ground-truth labels are unavailable in the target domain, this prior cannot be computed directly. Instead, we estimate it with simple region statistics from anatomical prior knowledge, which we denote as $\tau_e(t,k)$. Furthermore, the class-ratio of the segmentation network output prediction can be computed as follows: $\widehat{\tau}(t,k,\theta) = \frac{1}{|\Omega_t|}\sum_{i\in\Omega_t} p_t^k(i,\theta)$. We regularize the entropy in Eq. (3.2) with a Kullback-Leibler (KL) divergence matching these two class-ratios. Thus, our method minimizes the following overall loss during the training of the

adaptation phase:

$$\min_{\theta} \sum_{t} \frac{1}{|\Omega_t|} \sum_{i \in \Omega_t} \ell_{ent}(\mathbf{p}_t(i,\theta)) + \mathrm{KL}(\widehat{\tau}(t,\theta,\cdot), \tau_e(t,\cdot)) \tag{3.3}$$

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

Clearly, minimizing our overall loss in Eq. (3.3) during adaptation does not use the source images and labels. In the following, we discuss an interesting link between our loss in Eq. (3.3) and maximizing the mutual information between the target images and their network predictions. Figure 3.2 shows the overview of the proposed framework.

3.2.1 Link to mutual-information maximization

Notice that the terms of the KL penalty in Eq. (3.3) are inverted compared to our initial formulation (*AdaEnt*), which we provided in the conference version of this work Bateson *et al.* (2020); see Eq. (3.10). Besides the empirical motivation (as it will be shown in the experimental section hereafter), this is first and foremost motivated by theoretical results in information theory, as we link below Eq. (3.3) to maximizing the mutual information between the input images and their latent label predictions. The full proof is derived in I.

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

$$\begin{aligned} \mathcal{I}(X;Y) &= H(Y) - H(Y \mid X) \\ &= -\mathbb{E}_Y\left[\log \mathbb{E}_X[p(Y \mid X)]\right] + \mathbb{E}_{X,Y}[\log p(Y \mid X)] \end{aligned} \tag{3.4}$$

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

We denote P_t the $K \times |\Omega_t|$ softmax prediction mask, i.e. matrix whose columns are the vectors of network outputs $\mathbf{p}_t(i,\theta), i \in \Omega_t$. Given the classical interpretation of the softmax predictions as probabilities: $p_t^k(i,\theta) = p(y_t^k(i) = 1 \mid I_t, \theta)$, the empirical class-ratio distribution is an

estimate of the marginal distribution of P_t: $\hat{\tau}(t,\theta,\cdot) = \frac{1}{|\Omega_t|}\sum_{i\in\Omega_t}\mathbf{p}_t(i,\theta) = \mathbb{E}_{I_t}[p(P_t \mid I_t)]$. Therefore the empirical estimate of the mutual information between the images I_t and their softmax predictions, $P_t, t = 1,\ldots,T$, can be expressed as[9]:

$$\mathcal{I}_\theta = \frac{1}{T}\sum_t \underbrace{H\{\hat{\tau}(t,\theta,\cdot)\}}_{-\mathbb{E}_{P_t}\left[\log \mathbb{E}_{I_t}[p(P_t|I_t)]\right]} \underbrace{- \frac{1}{|\Omega_t|}\sum_{i\in\Omega_t}\ell_{ent}(\mathbf{p}_t(i,\theta))}_{-\mathbb{E}_{I_t,P_t}[\log p(P_t|I_t)]} \tag{3.5}$$

In the different context of discriminative clustering, Krause *et al.* (2010) draw a connection between maximizing the empirical estimate of the mutual information, as in Eq. (3.5), and a generalization of the mutual information based on the KL divergence, as in Eq. (3.3). Indeed, note that the following basic identity holds:

$$H\{\hat{\tau}(t,\theta,\cdot)\} \stackrel{c}{=} -KL\{\hat{\tau}(t,\theta,\cdot), U\} \tag{3.6}$$

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

$$\max_\theta \sum_t -KL\{\hat{\tau}(t,\theta,\cdot), \tau_e(t,\cdot)\} - \sum_t \frac{1}{|\Omega_t|}\sum_{i\in\Omega_t}\ell_{ent}(\mathbf{p}_t(i,\theta) \tag{3.7}$$

which is equivalent to Eq. (3.3). Maximizing the mutual information between the images I_t and their softmax predictions $p_t(\theta)$ is a principled approach in unsupervised problems, such as unsupervised discriminative clustering Jabi *et al.* (2021); Krause *et al.* (2010), further motivating our formulation, which we denote *AdaMI* in the following.

[9] See details of proof in I

3.2.2 Choosing the penalty function

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

$$\begin{aligned} \mathcal{L}_1 &= \mathrm{KL}(\tau_e(t,\cdot), \widehat{\tau}(t,\theta,\cdot)) \\ \mathcal{L}_2 &= \mathrm{KL}(\widehat{\tau}(t,\theta,\cdot), \tau_e(t,\cdot)) \end{aligned} \tag{3.8}$$

Figure 3.3 shows the profile of these two regularizers as functions of the class-ratio for a binary-segmentation case, with a target foreground class-ratio set to 0.5. We see that $\mathcal{L}_2$ may be a better choice than $\mathcal{L}_1$ when the initial predictions of the network are extremely imbalanced. Indeed, note the gradient properties and stability at the vicinity of 0, i.e., when the predicted foreground class-ratio $\tau(t,1)$ is close to 0. We see that both first and second derivatives of the regularizer are unbounded for $\mathcal{L}_1$, but bounded and constant for $\mathcal{L}_2$. Our experiments confirm the superiority of the $\mathcal{L}_2$ regularizer, in terms of training stability and quantitative performance.

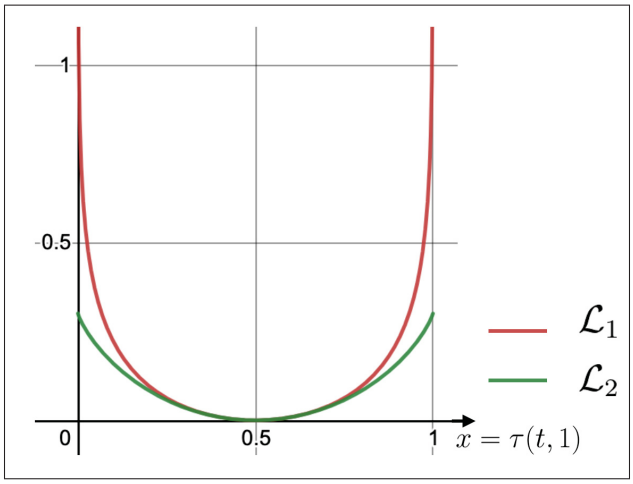


Figure 3.3 Comparison of two class-prior losses in the scenario $K = 2$, with the ground-truth class-ratio set to $\tau_{GT}(t,1) = 0.5$: The plots illustrate better gradient dynamics of $\mathcal{L}_2$ at the vicinity of a class-ratio $\tau(t,1) = 0$.

3.2.2.1 Estimating the class-ratio prior from anatomical knowledge

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

For each 2D target image I_t and each structure k, the class-ratio used for adapting the segmentation network with Eq. (3.3) is obtained by adding weak supervision in the form of image-level tag information:

$$\tau_e(t,k) = \begin{cases} \bar{\tau}_k & \text{if region } k \text{ is within image } t. \\ 0 & \text{otherwise,} \end{cases} \tag{3.9}$$

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

3.3 Experiments and Results

3.3.1 Experimental Settings

3.3.1.1 Data sets

3.3.1.1.1 IVDM3Seg

The proposed SFDA method is first evaluated on the dataset from the MICCAI 2018 IVDM3Seg Challenge[10], consisting of 16 3D multi-modality MRI data sets, collected from 8 subjects at two different stages to study inter-vertebral disc (IVD) degeneration. The scans were generated by a Dixon protocol with a 1.5 T Siemens MRI scanner, producing four aligned modalities. Scans are acquired in sagittal direction. Each volume has an anisotropic resolution of $2 \times 1.25 \times 1.25$ mm/vx. The corresponding manual segmentations of the IVDs are also available. In our experiments, we set the water modality (Wat) as the source and the in-phase (IP) modality as the target domain. Therefore, in this setting, the source and target modalities are acquired from the same patient. From this dataset, 12 scans are used for training, one for validation, and the remaining 3 scans for testing. Images are normalized to zero mean and unit variance. Then, we performed a data augmentation based on affine transformations. The setting is binary segmentation (K=2).

3.3.1.1.2 NCI-ISBI13

We employ prostate T2-weighted MRIs from 2 different data sources with distribution shifts from the NCI-ISBI13 dataset, with their corresponding manual segmentations of the prostate region. The source dataset consists of 30 volumes from Radboud University Nijmegen Medical Centre, generated with a 3 T Siemens scanner. Each source volume has an anisotropic resolution of $0.4 \times 0.4 \times 3$ mm/vx. The target dataset consists of 30 volumes from Boston Medical Center generated with a 1.5 T Philips Achieva. Each target volume has an anisotropic resolution

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

of 0.6-0.625 $\times$ 0.6-0.625 $\times$ 3.6-4 mm/vx. We use the publicly available pre-processed data provided by Liu, Dou & Heng (2020), which resized each sample to 384 $\times$ 384 in axial plane, normalized it to zero mean and unit variance. We employed data augmentation based on affine transformations. We use 19 scans for training, one for validation, and the remaining 10 scans for testing.

3.3.1.1.3 MMWHS

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

3.3.1.2 Benchmark Methods

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

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

$$\sum_t \frac{1}{|\Omega_t|} \sum_{i \in \Omega_t} \ell_{ent}(\mathbf{p}_t(i,\theta)) + \lambda \mathrm{KL}(\tau_e(t,\cdot), \widehat{\tau}(t,\theta,\cdot)) \tag{3.10}$$

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

$$\mathcal{L}_s(\theta, \Omega_s) + \frac{\lambda}{T} \sum_{t=1}^{T} [\tau_e(t,\cdot) - \widehat{\tau}_t(t,\theta,\cdot)]^2 \tag{3.11}$$

Curriculum Domain Adaptation: We denote *AdaSource* the method adopted in Zhang *et al.* (2020a):

$$\mathcal{L}_s(\theta, \Omega_s) + \frac{\lambda}{T} \sum_{t=1}^{T} KL(\tau_e(t,\cdot), \widehat{\tau}_t(t,\theta,\cdot)) \tag{3.12}$$

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

$$\mathcal{L}_s(\theta, \Omega_s) - \frac{\lambda}{T} \sum_{t=1}^{T} \sum_{i \in \Omega_T} \log\left(D\left(p_t(i,\theta)\right)^{(1)}\right) \tag{3.13}$$

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

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

We also compared to the following two source-free domain adaptation methods. The first is TTA Karani *et al.* (2021), which trains an auxiliary denoising network on the source, then applies it to the noisy segmentations in the target. The second is Tent Wang *et al.* (2021), which uses

a simple entropy minimization, similarly to Eq. (3.2). Importantly, for both methods, instead of optimizing the whole segmentation network, only the normalization statistics and affine parameters of the network are updated, while the rest of the parameters are frozen.

A model trained on the source domain only using Eq. (3.1), *NoAdap*, is used as a lower bound. A model trained with the supervised cross-entropy loss on the target domain, referred to as $Oracle$, serves as an upper bound.

Finally, for the cardiac application, we also present benchmark results obtained in previous DA works (Bian *et al.* (2020); Dou *et al.* (2019)), which we directly report in Table 3.2. The methods using AdaNet as the backbone were implemented in Dou *et al.* (2019), those with DeepLabV2 were implemented in Bian *et al.* (2020).

3.3.1.3 Evaluating robustness to class-ratio prior imprecision

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

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

$$\tau_{GT}(t,k) = \frac{1}{|\Omega_t|} \sum_{i \in \Omega_t} y_t^k(i) \tag{3.14}$$

This setting is hereafter referred as $AdaMI_{\tau_{GT}}$. This is followed by evaluating the robustness of our benchmarked method to a varying imprecision of the prior knowledge on the class-ratio prior,

i.e., varying the size estimates of the segmented regions.For each image t and each structure k of the target domain (except the background), we use the following error on the class-ratio:

$$\tau(t,k) = (1 \pm \delta)\tau_e(t,k) \tag{3.15}$$

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

3.3.1.4 Ablation study on target training dataset size

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

3.3.1.5 Ablation study on the weak annotations in the target training dataset

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

$$I_t \text{ is } \begin{cases} \text{selected, with } \tau_e(t,k) = \bar{\tau}_k & \text{if } \widehat{\tau}(t,\tilde{\theta},k) > \frac{1}{4}\bar{\tau}_k. \\ \text{selected, with } \tau_e(t,k) = 0 & \text{if } \widehat{\tau}(t,\tilde{\theta},k) = 0. \\ \text{discarded otherwise} \end{cases} \tag{3.16}$$

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

3.3.1.6 Training and implementation details

For all the methods, we employed UNet Ronneberger *et al.* (2015), a widely used segmentation network due to its simplicity. The architecture used is the same one as for the original UNet paper. We use a 2D implementation for all applications. In the source training phase, a model is trained on the source data only with Eq. (3.1) for 150 epochs, a learning rate of 5×10^{-4}, and a learning rate decay of 0.9 every 20 epochs. The final model is used as initialization to the adaptation phase. In this phase, the model is adapted with Eq. (3.3), trained with the Adam optimizer Kingma & Ba (2014), for 150 epochs. For all applications, the initial learning rate is 1×10^{-6}, the weight decay is 10^{-3}, and the batch size is 24. The learning rate decay is 0.7 for the heart and prostate applications, and 0.2 for the spine one. It is applied every 20 epochs. For all methods, we pick the final model as the one achieving the best validation score. The weights from Eq. (3.2) are calculated as: $\nu_k = \frac{\bar{\tau_k}^{-1}}{\sum_k \bar{\tau_k}^{-1}}$.

3.3.1.7 Evaluation metrics

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

3.3.2 Quantitative results

The quantitative performances of the different methods are presented in Table 3.1 for the spine and prostate images, and in Table 3.2 for the cardiac images.

Table 3.1 Performance comparison of the proposed formulation with different domain adaptation methods for spine (IVDM3Seg dataset, left) and prostate (NCI-ISBI13 dataset, right) segmentation, in terms of DSC (%) and ASD (vox)

Method	**Source Free**	**Target Tags**	**Spine IVDs**		**Prostate**	
			DSC	**ASD**	**DSC**	**ASD**
NoAdap (lower bound)	✓	×	68.5	2.15	67.2	10.59
Oracle (upper bound)	✓	✓	87.5	0.38	88.4	1.81
AdaptSegNet Tsai *et al.* (2018)	×	×	**82.4**	**0.50**	**83.1**	**2.43**
AdaSource Zhang *et al.* (2020a)	×	✓	75.9	0.99	76.3	3.93
CDA Bateson *et al.* (2021)	×	✓	75.7	0.86	77.9	3.28
TTA Karani *et al.* (2021)	✓	×	69.7	1.65	73.2	3.80
Tent Wang *et al.* (2021)	✓	×	68.8	1.84	68.7	5.87
Prior AdaEnt Bateson *et al.* (2020)	✓	✓	72.9	1.54	77.8	4.10
AdaMI (Ours)	✓	✓	74.2	1.17	79.5	3.92

3.3.2.1 No Adaptation

First, we see that the models trained with full supervision on the source domain suffer from a drop in performance when used in a different target domain without any adaptation. In Fig. 3.4(c), it can be verified that the *NoAdap* is in an under-segmentation regime, with the predicted sizes of structures well below their true sizes. This validates that the predictions are biased towards the dominant class, which is the background here.

3.3.2.2 With Adaptation

All models that use adaptation yield a substantial improvement over the lower baseline. For instance, on spine images, our model *AdaMI* reaches a Dice score (DSC) of 74.2%, representing 90% of the best-performing adaptation method, *AdapSegNet* Tsai *et al.* (2018), which used the source data during adaptation. *AdaMI* yields a 1.17 ASD, which corresponds to an improvement

by a multiplicative factor of 1.8 compared to the value for *NoAdap* (2.15 ASD). On prostate images, *AdaMI* reaches 79.5% DSC, 95% of the top performance *AdapSegNet*. An ASD of 3.92 is obtained, an improvement by a multiplicative factor of 3 compared to the value for *NoAdap* (10.59 ASD). Surprisingly, on cardiac images, where the domain shift is higher, *AdaMI* ranks second out of sixteen other adaptation techniques in terms of average DSC across cardiac structures, outperformed only by the recent method in Bian *et al.* (2020), a substantially more complex adaptation framework. Note that the quantitative results are not directly comparable between all models, since the backbone networks differ (see Table 3.2). These results show that having access to more information on source data does not necessarily help for the adaptation task. Finally, on all three applications, *AdaMI* outperforms the two other source-free domain adaptation methods. Specifically, *TTA* yields a smaller improvement than *AdaMI* on the spine and the prostate applications, and fails on the more difficult heart one. *Tent* only yields a small improvement in terms of Dice on all three applications.

3.3.2.3 AdaMI versus AdaEnt

The Dice scores (DSC) of our proposed *AdaMI* reach 85% of *Oracle*'s performance on spine images, 90% of its performance on prostate images, and 85% on cardiac images. This validates the efficiency of using a class-ratio prior matching with a KL divergence to prevent under-segmentation. Comparing *AdaMI* and *AdaEnt*, we see that on all three applications, *AdaMI* outperforms *AdaEnt* and shows better convergence properties (see Fig. 3.4 (b)). Moreover, in Fig. 3.4 (a), we can observe that *AdaEnt* reaches rapidly its highest validation DSC (first 20 epochs) before slowly decaying. Fig. 3.4 (c) shows that the mean predicted size of structures jumps instantly from 50% below to 15% above the mean ground-truth sizes before stagnating. On the contrary, the performance of *AdaMI* improves steadily and the sizes of predicted structures grow progressively. This suggests that the inversion of the terms in the KL divergence in *AdaMI*, such as in Eq. (3.3), does help the learning process in domain adaptation, when compared to the original KL divergence in *AdaEnt* (see Section 3.2.2). Finally, the ASD values confirm the trend across the different models on cardiac images. Improvement over the lower baseline model (14.6

voxels) is substantial for *AdaEnt* (8.2 voxels), and even greater for *AdaMI* (5.6 voxels), with the greatest improvement occurring for AA and LA structures.

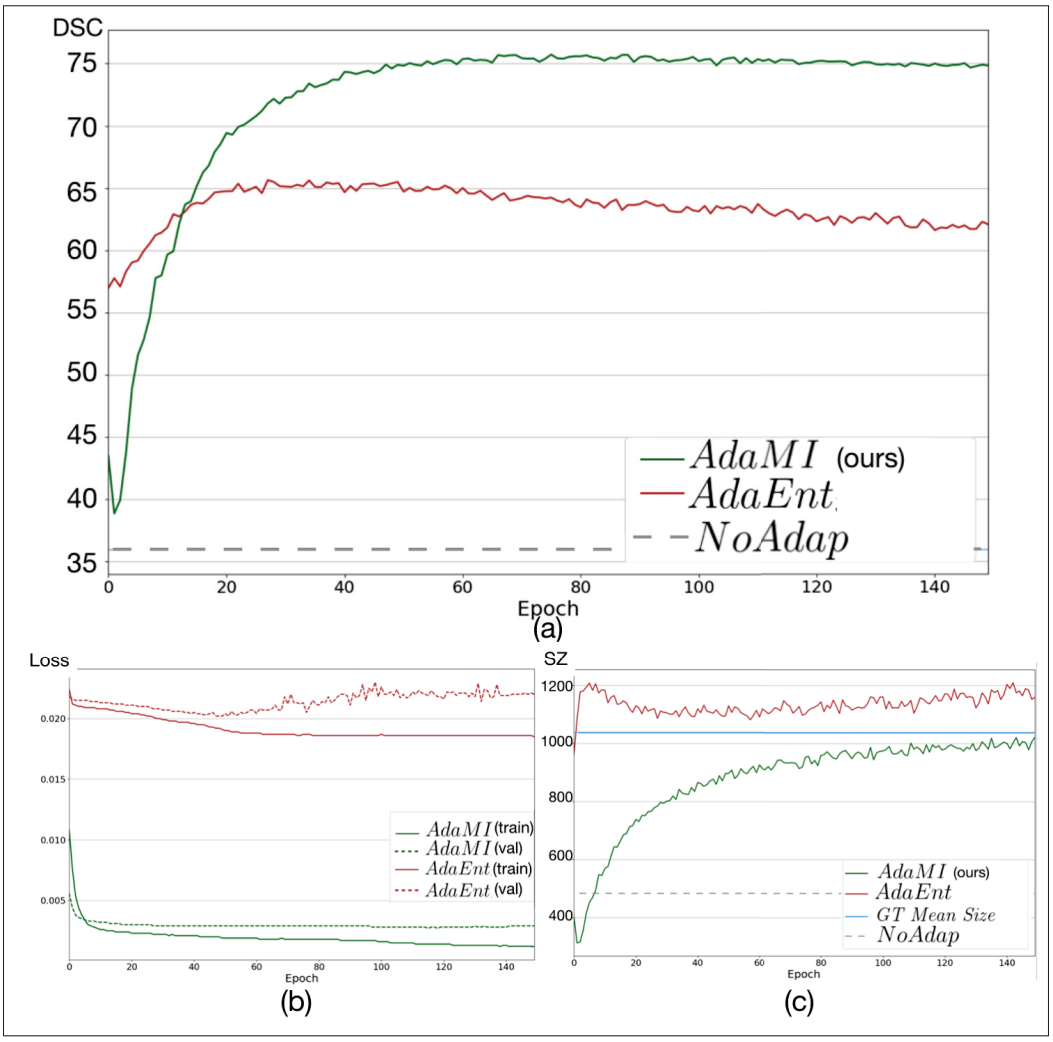


Figure 3.4 Quantitative performance: (a) Evolution of DSC (%) and (b) Learning Curves and (c) mean ground truth sizes and predicted sizes (px) of cardiac structures segmentation masks over training epochs on target images from the validation set. Comparison of the proposed model *AdaMI*, and our previous *AdaEnt*.

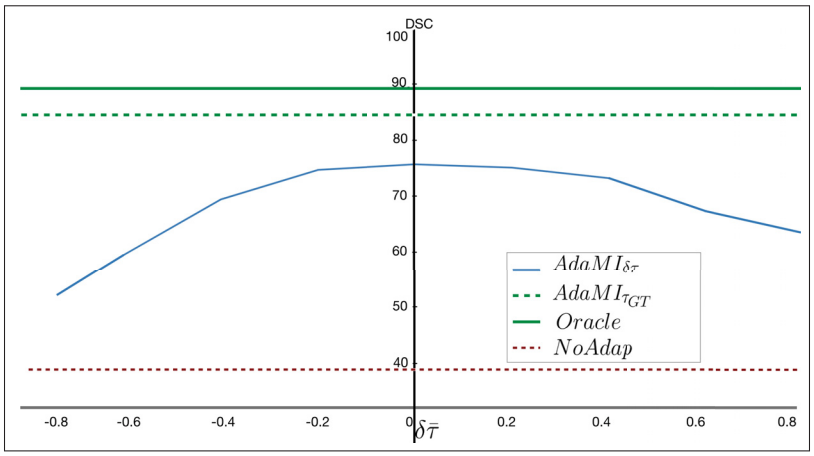


Figure 3.5 Robustness performance: DSC (%) versus enforced relative size error in the class-ratio prior $\delta\bar{\tau}$ for each structure for cardiac segmentation, showing robustness to imprecision in the prior. The DSC performance of the upper bounds *Oracle*, $AdaMI_{\tau_{GT}}$ and lower bound *NoAdap* are also indicated.

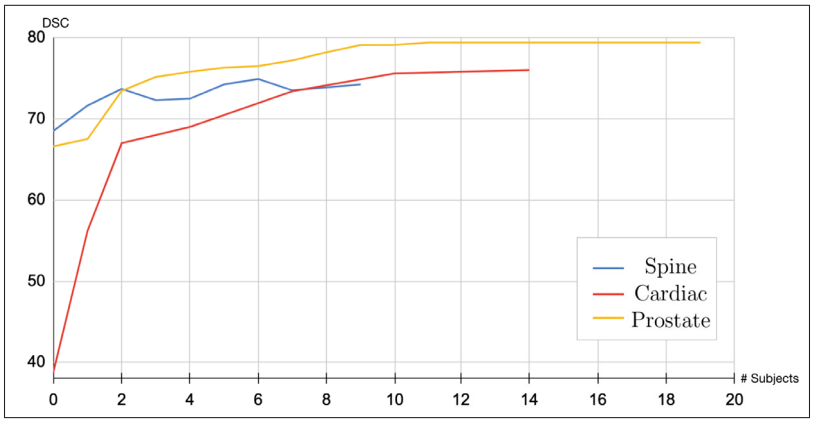


Figure 3.6 Ablation performance: DSC (%) in target test set versus number of subject in the target training dataset for each application, showing the data efficiency of our method.

3.3.3 Ablation study on class-ratio precision

We also investigate the impact of imprecision in the target domain class-ratio prior on the quality of SFDA models. To this end, we validate a range of values in the estimations of class-ratios,

Table 3.2 Performance comparison of the proposed formulation with different domain adaptation methods for cardiac segmentation, in terms of DSC (mean) and ASD (mean)

Methods	**Source Free**	**Target Tags**	**Backbone**	DSC (%)					ASD (vox)				
				AA	**LA**	**LV**	**Myo**	**Mean**	**AA**	**LA**	**LV**	**Myo**	**Mean**
NoAdap (lower bound)	✓	×	UNet	49.8	62.0	21.1	22.1	38.8	19.8	13.0	13.3	12.4	14.6
Oracle (upper bound)	✓	✓		91.9	88.3	91.0	85.8	89.2	3.1	3.4	3.6	2.2	3.0
AdaSource Zhang *et al.* (2020a)	×	✓		79.0	77.9	64.4	61.3	70.7	6.5	7.6	7.2	9.1	7.6
CDA Bateson *et al.* (2021)	×	✓		77.3	72.8	73.7	61.9	71.4	**4.1**	6.3	6.6	**6.6**	5.9
TTA Karani *et al.* (2021)	✓	×		59.8	26.4	32.3	44.4	40.7	15.1	11.7	13.6	11.3	12.9
Tent Wang *et al.* (2021)	✓	×		55.4	33.4	63.0	41.1	48.2	18.0	8.7	8.1	10.1	11.2
Prior AdaEnt Bateson *et al.* (2020)	✓	✓		75.5	71.2	59.4	56.4	65.6	8.5	7.1	8.4	8.6	8.2
AdaMI (Ours)	✓	✓		**83.1**	**78.2**	**74.5**	**66.8**	**75.7**	5.6	**4.2**	**5.7**	6.9	**5.6**
AdaptSegNet Tsai *et al.* (2018)	×	×	DeepLabV2	65.4	80.6	81.4	69.3	74.2	8.1	5.3	4.0	3.6	5.2
BDL Li, Yuan & Vasconcelos (2019)	×	×		67.1	80.6	82.7	62.1	73.1	12.0	7.0	3.5	4.2	6.7
CLAN Luo, Zheng, Guan, Yu & Yang (2019)	×	×		63.8	79.9	**84.4**	66.8	73.7	9.1	5.3	**3.4**	**3.5**	5.3
DISE Chang, Wang, Peng & Chiu (2019a)	×	×		71.8	82.2	83.7	60.8	74.6	6.7	4.7	3.8	7.7	5.7
SynSeg-Net Huo *et al.* (2019)	×	×		71.6	69.0	51.6	40.8	58.2	11.7	7.8	7.0	9.2	8.9
UADA Bian *et al.* (2020)	×	×		**84.1**	**88.3**	84.3	**71.4**	**82.1**	**3.9**	**3.5**	3.8	**3.7**	**3.7**
CyCADA Hoffman *et al.* (2018)	×	×	AdaNet	72.9	77.0	62.4	45.3	64.4	9.6	8.0	9.6	10.5	9.4
SIFA Chen *et al.* (2020b)	×	×		**81.3**	**79.5**	**73.8**	**61.6**	**74.1**	**7.9**	**6.2**	**5.5**	**8.5**	**7.0**
PnP-AdaNet Dou *et al.* (2019)	×	×		74.0	68.9	61.9	50.8	63.9	12.8	6.3	17.4	14.7	12.8
CycleGAN Zhu *et al.* (2017)	×	×		73.8	75.7	52.3	28.7	57.6	11.5	13.6	9.2	8.8	10.8
DANN Ganin *et al.* (2016)	×	×		39.0	45.1	28.3	25.7	34.5	16.2	9.2	12.1	10.1	11.9
ADDA Tzeng *et al.* (2017)	×	×		47.6	60.9	11.2	29.2	37.2	13.8	10.2	NA	13.4	NA
Overall ranking of AdaMI (#/16)				2	7	6	3	**2**	3	2	7	6	**4**

as explained in Sec. 3.3.1.3. The results are reported for cardiac images in Fig. 3.5. First, in the ideal situation where the precise class-ratios are known, $AdaMI_{\tau_{GT}}$ reaches 84.5% DSC, representing 95% of the upper baseline, the *Oracle*. Then, we can see that our proposed method *AdaMI* is robust to large ranges of imprecision in class-ratio estimates. Indeed, a difference of ± 20% (resp. ± 40%) with our prior estimation in Sec. 3.2.2.1 only degrades the DSC by up to 1% (resp. 6%). Moreover, we see that an overestimation of the structure sizes leads to a better overall DSC than an underestimation, highlighting the well-known bias of Dice towards over-segmentation.

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

3.3.4 Ablation study on the size of the target training dataset

We also investigate how much weakly-labeled target training data is necessary for our SFDA model to achieve adaptation. To this end, we experiment with a varying number of subjects in the target training dataset. The results are reported in Fig. 3.6. We can see that our proposed method *AdaMI* is robust to large diminution of target dataset size. Indeed, with only 2 subjects, *AdaMI* is on par with most state-of-the art methods, reaching 67% DSC for the cardiac application, 74% DSC for the spine, and 73% for the prostate.

Table 3.3 Performance of the proposed formulation obtained when removing the weak image-level annotations

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

3.3.5 Ablation study on the weak annotations in the target training dataset

Finally, we investigate the more general scenario where images are fully unsupervised in the target domain. Particularly, we removed the target image tags for the adaptation phase as explained in Section 3.2.2.1. Results from this study are reported in Table 3.3. As expected, having image-level tag information helps all the models, which can be observed from the performance degradation compared to results in Table 3.1 and 3.2. Indeed, the class-ratio estimation degrades without the image tag, and as a result, models using a class-ratio prior to guide adaptation also see their performance decrease. However, for the spine and the prostate application, the quantitative performance (73.7% DSC and 71.8% DSC respectively) remains well above the baseline, on par with most state-of-the art domain adaptation models. The removing of image-level Tags is more difficult for the heart application, as it is multi-class and has a big domain shift. However, results (58.0% DSC) stayed well above both the baseline and the two other SFDA methods, *Tent* and *TTA*.

3.3.6 Qualitative results

Qualitative segmentations and the corresponding entropy maps are shown for spine images in Fig. 3.7, for prostate images in Fig. 3.8, and for cardiac ones in Fig. 3.9. Without adaptation, the predictions of the network are either uncertain, as revealed by the high activation in the entropy maps of predictions (see top two lines in Fig. 3.9); or severely biased towards the dominant class, i.e. the background. This bias produces under-segmented or completely undetected structures (see the top four rows in Fig. 3.9). In all cases, the output segmentation masks are noisy, with very irregular edges. Benchmark adaptation models *CDA* and *AdaSource* are able to recover the structures in most examples. However, they display high uncertainty in the predictions, especially *CDA*. Interestingly, for some difficult slices, the segmentation results produced by our proposed SFDA model matches better with the ground-truth. For spine and prostate images, such examples are displayed in bottom two rows in Fig. 3.8. For cardiac images, the whole AA structure is better recovered (see middle two rows in Fig. 3.9), and the shapes and the boundary between the MYO and the LV structure are improved. Notably, in all applications, the entropy maps produced by *AdaMI* only show high activations along the borders of the predicted structures. These visual results further confirm the remarkable ability of *AdaMI* to produce accurate predictions with high confidence over existing approaches.

3.4 Discussion

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

Source-Free Domain Adaptation: Surprisingly, even though our model does not access the source data in the adaptation phase, it yields comparable or better performance than many state-of-the art adaptation approaches that do rely on the source data. It also outperforms two

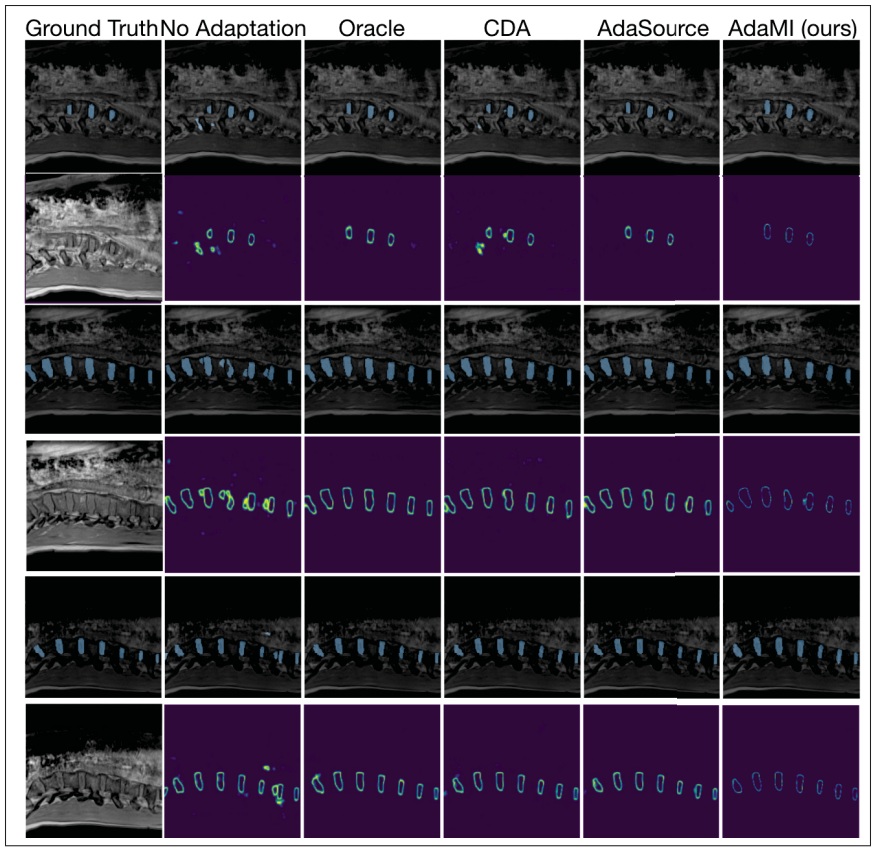


Figure 3.7 Qualitative performance on spine MRI images: examples of the segmentations achieved by our formulation (*AdaMI*), benchmark models in Bateson *et al.* (2021), Zhang *et al.* (2020a) and lower (*NoAdap*) and upper baselines (*Oracle*). First column shows an input slice and the corresponding semantic segmentation ground-truth. The other columns show segmentation results (top) along with prediction entropy maps produced by the different models (bottom).

very recent source-free domain adaptation approaches, Karani *et al.* (2021); Wang *et al.* (2021). These works have stressed on the need for limited flexibility at test time, by freezing most parameters in the network, and adapting only the normalization and affine ones. Yet, in our three applications, we have found our proposed method, where the entire network is adapted, to be more efficient. Furthermore, our principled solution to source-free domain adaptation minimizes the uncertainty of the target domain predictions while preventing trivial solutions of single-class outputs via a KL regularizer that encourages target class-ratio (i.e region proportions). Using

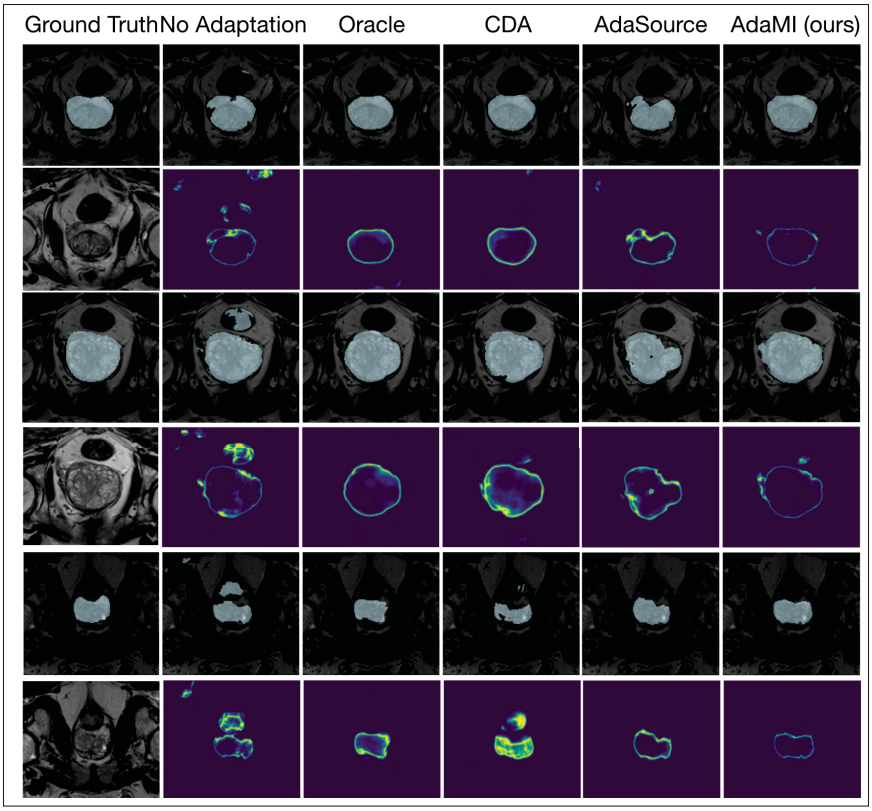


Figure 3.8 Qualitative performance on prostate MRI images: examples of the segmentations achieved by our formulation (*AdaMI*), benchmark models in Bateson *et al.* (2021), Zhang *et al.* (2020a) and lower (*NoAdap*) and upper baselines (*Oracle*).

entropy minimization in combination with this regularizer, our formulation reaches 85%, 90% and 85% of full supervision in spine, prostate, and cardiac images respectively. Our qualitative results demonstrate the ability of SFDA to produce accurate predictions with high confidence.

Robustness: Our experiments have further confirmed the robustness of *AdaMI* to substantial prior imprecision, and that having a coarse knowledge of the target region proportions can be enough to guide adaptation. In our implementation, a class-ratio prior is derived from readily available anatomical reference values. This anatomical knowledge is combined with image-level tags to produce a very coarse yet effective estimation of target class-ratios. This finding has great potential value in the medical domain, as prior anatomical knowledge is commonly available, due to conventions in patient position and anatomical similarity El Jurdi *et al.* (2021). We have,

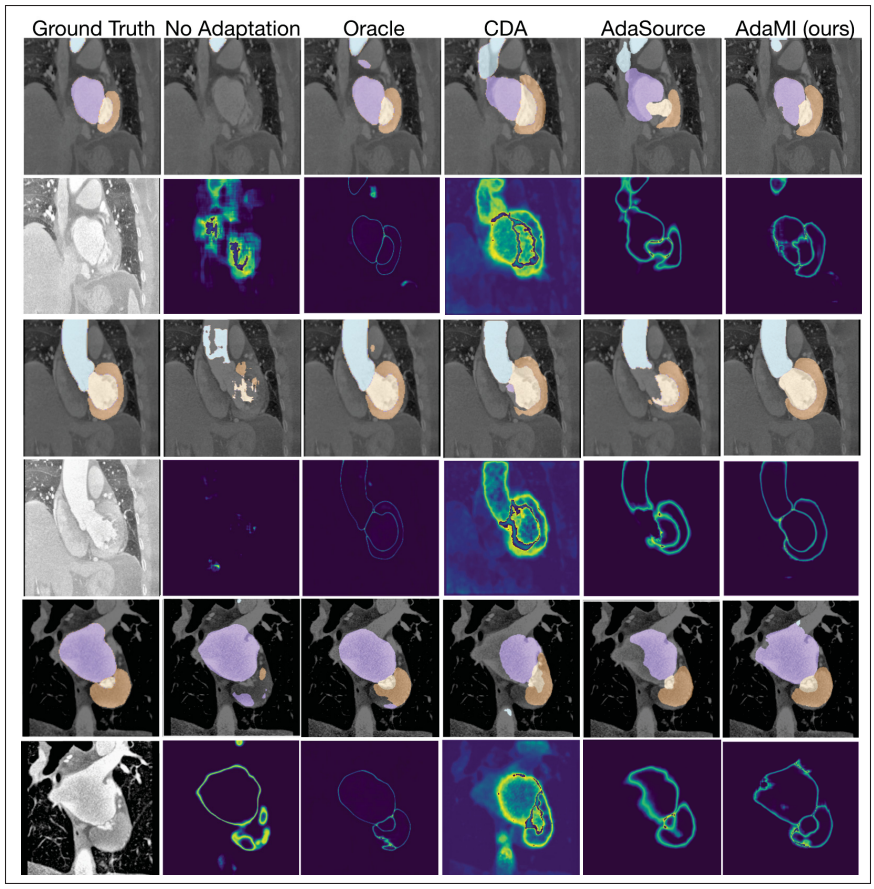


Figure 3.9 Qualitative performance on cardiac CT images: examples of the segmentations achieved by our formulation (*AdaMI*), benchmark models in Bateson *et al.* (2021), Zhang *et al.* (2020a) and lower (*NoAdap*) and upper baselines (*Oracle*). The cardiac structures of MYO, LA, LV and AA are depicted in brown, purple, yellow and blue, respectively

therefore, proposed an effective method to integrate such domain-invariant knowledge, with straightforward extensions in many medical applications. Moreover, the method seems robust enough to adapt to cohorts with possible anatomical variability, i.e. a large shift of class-ratio distributions compared to anatomical reference values (e.g. population-wise differences). Indeed, we show in Table that our model large ambiguities (±60%) on these class-ratios distributions only degrade the performance by up to 15%.

We have also shown that, in the ideal setting when a very precise prior is known, the performance of *AdaMI* is close to full supervision. This suggests that *AdaMI* is able to approach the *"optimal"*

segmentation given the amount of prior imprecision. This finding is in line with the recent work of Kervadec *et al.* (2021), which shows that using a few global shape descriptors as supervision enables performances close to a full pixel-wise supervision. In fact, the class-ratio used in *AdaMI* is based on zero-order shape moments.

We have also demonstrated the superiority of *AdaMI* when compared to our previous *AdaEnt*, which regularizes the class-ratio priors with a steeper loss Bateson *et al.* (2020). Indeed, *AdaMI* is able to prevent the under-segmentation regime observed without adaptation, while avoiding the fast convergence to local minima observed with *AdaEnt*. Although convergence and stability are well-known challenges for unsupervised and weakly supervised deep domain adaptation methods, *AdaMI* shows remarkable training stability. On cross-modality spine MRI and cross-site prostate MRI, our method has shown performances on par with several adaptation models that necessitate both the source and target data, such as Bateson *et al.* (2021); Zhang *et al.* (2020a). Surprisingly, for the adaptation of MRI-to-CT cardiac images, our model outperforms several recent state-of-the-art adaptation models, such as Bateson *et al.* (2021); Chen *et al.* (2020b); Ganin *et al.* (2016); Hoffman *et al.* (2018); Tsai *et al.* (2018); Tzeng *et al.* (2017); Zhang *et al.* (2020a); Zhu *et al.* (2017). This is confirmed qualitatively by our experiments, where the structures of interests are well predicted in all the three applications. In some cases, the segmentation masks are even improved when compared to benchmark adaptation models, despite the lack of source data. These results, therefore, suggest that having access to the source data may not be necessary for domain adaptation.

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

Limitations: A limitation of our work is the need for an image-level annotation, compared to fully unsupervised domain adaptation methods. Such annotation for each slice of each

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

3.5 Conclusion

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

computational burden of the domain adaptation, the optimization complexity, and can be used with any segmentation network architecture.

CHAPTER 4

SINGLE-SUBJECT TEST-TIME ADAPTATION WITH SHAPE MOMENTS FOR IMAGE SEGMENTATION

Mathilde Bateson[1], Hervé Lombaert[1], Ismail Ben Ayed[1]

École de Technologie Supérieure, 1100 Notre-Dame West, Montreal, Quebec, Canada H3C 1K3

Paper submitted to *Medical Image Analysis* (MEDIA).

Presentation

This chapter presents the article "Single-Subject Test-Time Adaptation with Shape Moments for Image Segmentation" submitted to Medical Image Analysis, ongoing major revision. An initial article was published in the MICCAI conference, 2022, presented in Singapour. The objective of this article is to propose a single-subject test-time adaptation framework from segmentation. We explore the potential of combining entropy minimization with various constraints in the form of shape moments, to guide domain adaptation towards plausible solutions. In particular, we exploit the size, centroid, and distance-to-centroid of anatomical structures through penalty constraints in our overall loss function.

4.1 Introduction

Deep learning models have achieved many breakthroughs, approaching human-level performances in various natural and medical-imaging problems, when trained on large-scale labeled data Litjens *et al.* (2017). Nevertheless, these systems may fail when the distribution of the test images is different from those seen during training. For instance, this can be caused by variations in medical imaging modalities, such as training on MRI scans and testing on CT scans, by acquisition-related domain shifts, such as different protocols, vendors, machines or clinical sites, and by differences in subject populations. For semantic segmentation tasks, labelling images in each new target dataset is time-consuming, heavily burdening clinical applications, as it requires precious expert knowledge. To circumvent those impediments, methods that learn robust networks with less supervision have attracted substantial attention in medical imaging Cheplygina *et al.* (2019).

4.1.1 Domain Adaptation (DA)

Reduced supervision motivates *Domain Adaptation* (DA) methods: DA consists in adapting a model trained on an annotated source domain to a unseen target domain with no or minimal new annotation. Popular strategies for both classification and segmentation tasks include minimizing the discrepancy between source and target distributions in the feature or output spaces Tsai *et al.* (2018); integrating a domain-specific module in the network Dou *et al.* (2019); translating images from one domain to the other Zhu *et al.* (2017); or incorporating a domain-discriminator module and penalizing its success in the loss function Tzeng *et al.* (2017).

One of the key drawback of DA methods is that they require the source dataset to be present while training for each new target dataset. In medical applications, sharing datasets across different clinical sites is often impossible for privacy and regulatory reasons. As the source and target data may come from different institutions, there is a critical need to develop DA methods which do not assume access to the source data. Standard methods, such as Dou *et al.* (2019); Tsai *et al.* (2018); Tzeng *et al.* (2017); Zhu *et al.* (2017), do not comply with this restriction.

This has motivated *Source-Free Domain Adaptation* (SFDA) Bateson *et al.* (2020); Karani *et al.* (2021), a setting where the source data is unavailable during the training of the adaptation phase, neither in the form of source images nor ground-truth masks.

4.1.2 Test-Time Adaptation (TTA)

Evaluating DA or SFDA methods consists in: (i) performing the chosen adaptation strategy on a dedicated training set *Tr* from the target domain; (ii) choosing the final model by measuring metrics on a validation set *Tv* and (iii) measuring the generalization performance on an unseen test set *Te* in the target domain. However, emerging works in the field of *Test-Time Adaptation* (TTA) Wang *et al.* (2021) argue that it is more efficient to adapt directly to the specific subjects from the test set *Te*. Moreover, access to the target distribution may not be necessarily presumed in clinical applications, often leaving only a single subject available in the target domain at test time.

Amongst important considerations in TTA are the risks of over-adaptation of the model parameters, along with the choice of hyper-parameters, as there are no dedicated training and validation sets. Indeed, Boudiaf *et al.* (2022) shows that recent deep-network based TTA methods for image classification perform well only when tested in narrowly defined experimental setups, and may fail drastically otherwise. To address this, an emerging method is to update the network batch normalization scale and bias parameters only, while freezing the rest of the network trainable parameters. This was leveraged in Wang *et al.* (2021), which adapts a model to new target images through entropy minimization. Similarly, Liang *et al.* (2020) uses entropy minimization with a diversity regularizer, while Mummadi, Hutmacher, Rambach, Levinkov, Brox & Metzen (2021) further adds an input transformation module. Instead of updating the trainable parameters in batch normalization layers, another popular direction is to update the batchnorm statistics using the target data Hu *et al.* (2021b). Amongst alternative approaches in classification, Boudiaf *et al.* (2022) proposes Laplacian regularization, which only updates the probability outputs of the model but not its trainable parameters. Finally, Wang, Fink, Van Gool & Dai (2022) tackles continual TTA, where the target domain distribution

may change over time, and proposes to stochastically restore a small part of the weights to the source pre-trained weights in order to prevent catastrophic forgetting. A few recent related works investigated the TTA setting for semantic segmentation tasks: The work in Kundu, Kulkarni, Singh, Jampani & Babu (2021) used a prior-enforcing auto-encoder to discourage spatial irregularities; whereas Liu, Zhang & Wang (2021d) generates fake images to promote feature alignment, without access to the source data. In the context of medical imaging, Hu, Song, Gu, Luo, Chen, Chen et al. (2021a) adopts regularizers to improve segmentation contours and class diversity. Recently, Karani *et al.* (2021) trains an auxiliary denoising network on the source images, and applies it to the noisy segmentations of the target domain. A drawback of this method is its modification of the source training stage. Therefore, it cannot be used with an off-the-shelf segmentation model.

4.1.3 Domain Generalization

Broadly related to our work, Domain Generalization (DG) aims at improving out-of-distribution generalization performance on potentially unseen target domains. DG also foregoes the need to access a large set of samples from the target distribution, by learning a model from multiple source domains. As a first simple baseline approach, data augmentation during training Hendrycks, Basart, Mu, Kadavath, Wang, Dorundo et al. (2021) or testing Ashukha, Lyzhov, Molchanov & Vetrov (2020) has been shown to improve model robustness. Common recent DG strategies rely on domain discrepancy minimization Motiian, Piccirilli, Adjeroh & Doretto (2017), meta-learning to stimulate domain shifts Liu *et al.* (2020), and federated learning from multiple decentralized sources Liu, Chen, Qin, Dou & Heng (2021b). However, one fundamental limitation of DG methods is that they aim to improve the generalization capacity of a model by leveraging source domains, and lack *test-time adaptability*. Instead, our approach considers to focus on improving the performance of an existing pre-trained neural network during test-time by directly adapting it to a specific subject from the target domain data.

4.1.4 High-level priors on shape moments for image segmentation

Amongst models that learn a shape representation, active and statistical shape models have been successfully integrated with deep segmentation networks Bohlender, Oksuz & Mukhopadhyay (2021). They are either applied as pre-processing or post-processing steps, or used in multi-steps models. However, as they impose the shapes to be consistent with the ones observed in a training set, deviations from the training shapes may be inadequate if anatomies vary. This is particularly inconvenient in medical image segmentation where anatomical malformations should be detected rather than ignored. A practical alternative to building these complex models is to use high-level priors El Jurdi *et al.* (2021), which are simpler to obtain while allowing local deformations. Such priors may inform on the object shape, size, topology, or inter-region constraints. Examples include the star-shape prior in Veksler (2008), which has been successfully integrated in optic-disc Bai, Miri, Liu, Saha, Garvin & Wu (2014) and skin-lesion segmentation Mirikharaji & Hamarneh (2018), as well as region-interaction priors for segmenting objects while preserving topology Schmidt & Boykov (2012).

Shape moments

The high-level priors considered in our work are shape-moment constraints, which were investigated in classical interactive segmentation settings Klodt & Cremers (2011a), well before modern deep learning paradigms. In particular, low-order moments characterize the volume, centroid and covariance of the segmentation regions; they do not embed information on fine-grained shape characteristics. Low-order shape moments were used before in the general context of deep segmentation networks. For instance, the zero-order moment, which represents the volume (or size) of the target segmentation region, was investigated as a prior to guide the training of deep segmentation models under limited supervision, including the weakly Jia *et al.* (2017); Kervadec *et al.* (2019b) or semi-supervised settings Kervadec *et al.* (2019a); Zhou *et al.* (2019a). In domain adaptation works, various form of this size prior were used as a source of additional supervision in Bateson *et al.* (2021); Vu *et al.* (2019); Zhang *et al.* (2020a). Recently, the work in Kervadec *et al.* (2021) showed that, in the fully supervised segmentation setting,

supervision in the form of a few shape moments could achieve performances close to supervision with pixel-wise labels.

4.1.5 Contributions

We propose a simple formulation for source-free and single-subject test-time adaptation (TTA) of segmentation networks. During inference for a single test subject, we optimize a loss integrating shape-moment priors and the entropy of predictions with respect to the scale and bias parameters of the batch-normalization layers. Unlike the standard source-free domain adaptation (SFDA) setting, we perform test-time adaptation on each subject separately, and forgo the use of training and validation set from the target domain during adaptation. Our setting is most similar to the image classification work in Wang *et al.* (2021), which minimized a label-free entropy loss defined over test-time samples. Building on this entropy loss, we further guide segmentation adaptation with shape-moment priors on the target regions, and show their substantial effect on TTA performances.

We conduct extensive experiments and comparisons with state-of-the-art TTA, SFDA and DA methods on two segmentation tasks: MRI-to-CT adaptation for cardiac images, from MRI to CT, and cross-site adaptation adaptation for prostate MRI images. These experiments show the effectiveness of our shape-guided entropy minimization method, which significantly improves the segmentation performance when compared to existing TTA methods. Surprisingly, it also fares better than various state-of-the-art SFDA and DA methods, even though it is not trained on a source and additional target data during adaptation. Instead, we perform joint inference and adaptation on a single data point in the target domain. Our results question the usefulness of training on a target set during adaptation and point to the effectiveness of embedding shape-moment priors during inference on domain-shifted testing data.

A preliminary conference version of this work appears in MICCAI 2022 Bateson, Kervadec, Dolz, Lombaert & Ben Ayed (2022a). This journal version provides significant extensions: (1) we introduce a 3D version of the method, using volume-wise instead of the slice-wise shape

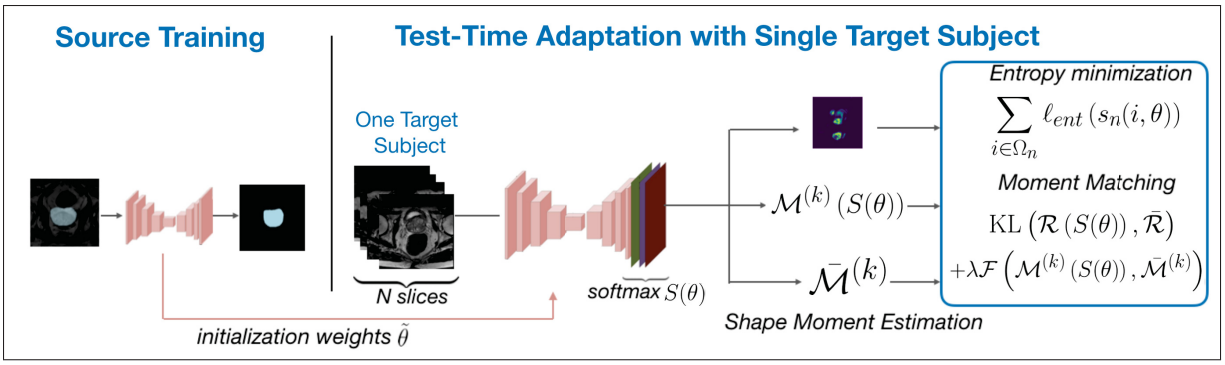


Figure 4.1 Overview of our framework for Test-Time Adaptation with Shape Moments: we leverage entropy minimization and shape-moment priors to adapt a segmentation network on a single subject at test-time.

information in Bateson *et al.* (2022a); (2) while Bateson *et al.* (2022a) required a weak annotation, we study here a fully unsupervised setting as well as different types/ levels of annotations; (3) we study the combination of multiple shape moments. We believe that our approach offers a great potential in multiple clinical settings, particularly when prior knowledge on the shape of the target structures is easily accessible. Our framework can be readily used for adapting a breadth of segmentation problems, with the code made publicly available[11].

4.2 Method

4.2.1 Definitions and Notations

Given an image $I : \Omega \subset \mathbb{R}^3 \rightarrow \mathbb{R}$, and its ground-truth K-class segmentation mask $\mathcal{Y}$, we denote the one-hot encoding for each voxel $i \in \Omega$ as a K-simplex vector $\mathbf{y}(i) = \left(y^{(1)}(i), \ldots, y^{(K)}(i)\right) \in \{0,1\}^K$. For a given class k, we also write the k-th component of $\mathcal{Y}$ in tensor form : $\mathbf{Y}^{(k)} = \left(y^{(k)}(i)\right)_{i\in\Omega}$.

For each voxel i, its coordinates in the 3D space are represented by the tuple $\left(u_{(i)}, v_{(i)}, w_{(i)}\right) \in \mathbb{R}^3$.

[11] https://github.com/mathilde-b/TTA

Shape moments and descriptors

Given an image I from a domain $\Omega \subset \mathbb{R}^3$ and its ground truth segmentation mask $\mathcal{Y}$, we introduce associated shape moments: each shape moment is parametrized by its orders $p, q, r \in \mathbb{N}^3$, and each order represents a different characteristic of the shape. For a given $p, q, r \in \mathbb{N}$ and class k, the 3D shape moments can be computed as follows [12] :

$$\mu_{p,q,r}\left(\mathbf{Y}^{(k)}\right) = \sum_{i \in \Omega} y^{(k)}(i) u_{(i)}^{p} v_{(i)}^{q} w_{(i)}^{r} \tag{4.1}$$

Central moments are derived from shape moments to guarantee translation invariance. They are computed as follows:

$$\bar{\mu}_{p,q,r}\left(\mathbf{Y}^{(k)}\right) = \sum_{i \in \Omega} y^{(k)}(i) \left(u_{(i)} - \bar{u}^{(k)}\right)^{p} \left(v_{(i)} - \bar{v}^{(k)}\right)^{q} \left(w_{(i)} - \bar{w}^{(k)}\right)^{r}. \tag{4.2}$$

where $\left(\bar{u}^{(k)}, \bar{v}^{(k)}, \bar{w}^{(k)}\right) = \left(\frac{\mu_{1,0,0}(y^{(k)})}{\mu_{0,0,0}(y^{(k)})}, \frac{\mu_{0,1,0}(y^{(k)})}{\mu_{0,0,0}(y^{(k)})}, \frac{\mu_{0,0,1}(y^{(k)})}{\mu_{0,0,0}(y^{(k)})}\right)$ are the components of the 3D centroid.

We compute these moments for each class $k \in \{1, \ldots, K\}$ and use the vectorized form onwards, e.g.:

$$\mu_{p,q,r}\left(\mathcal{Y}\right) = \left(\mu_{p,q,r}(\mathbf{Y}^{(1)}), \ldots, \mu_{p,q,r}(\mathbf{Y}^{(K)})\right)^{\top}. \tag{4.3}$$

Note that given a 3D volume, the shape moments can either be embedded in 2D, by computing them for each slice of the volume, or in 3D, by computing them on the whole volume. In the following, we present the method focusing on the 3D version, while we present results for both versions in Section 4.3.

[12] The moments of a 2D image can be computed similarly, simply removing the third component corresponding to the w coordinate.

4.2.2 Description of the method

4.2.2.1 Pre-training Phase

We consider a set of M source slices $I_m : \Omega_s \subset \mathbb{R}^2 \rightarrow \mathbb{R}$, $m = 1, \ldots, M$, and their ground-truth K-class segmentation mask $\mathcal{Y}_m$. The network is first trained on the source domain Ω_s only, by minimizing the cross-entropy loss with respect to network parameters θ:

$$\min_{\theta} \frac{1}{|\Omega_s|} \sum_{m=1}^{M} \ell \left(\mathbf{y}_m(i), \mathbf{s}_m(i, \theta) \right) \tag{4.4}$$

where $\mathbf{s}_m(i, \theta) = (s_m^{(1)}(i, \theta), \ldots, s_m^{(K)}(i, \theta)) \in [0, 1]^K$ denotes the predicted softmax probability for class $k \in \{1, \ldots, K\}$, which we denote $S_m \in [0, 1]^{K \times |\Omega|}$ in tensor form.

4.2.2.2 Shape moments on the target predictions

Given a subject $\mathcal{I} \subset \Omega_T$ in the target domain, we derive the shape moments from the network segmentation prediction $S(\theta)$. We build from the definitions in Section 4.2.1, replacing the ground truth $\mathcal{Y}$ by the softmax $S(\theta)$, to obtain the shape moments of the predictions: $\mu_{p,q,r}(S(\theta))$, with $(p, q, r) \in \mathbb{N}^3$. We then derive the shape descriptors $\mathcal{R}, \mathcal{C}, \mathcal{D}$ defined in Table 4.1 for $\Omega_T \in \mathbb{R}^3$, which respectively inform on the size, position, and compactness of a shape.

Table 4.1 Examples of 3D shape descriptors based on softmax predictions

Shape Descriptor	Definition
Class-Ratio	$\mathcal{R}(s) := \frac{\mu_{0,0,0}(s)}{\lvert\Omega_T\rvert}$
Centroid	$\mathcal{C}(s) := \left(\frac{\mu_{1,0,0}(s)}{\mu_{0,0,0}(s)}, \frac{\mu_{0,1,0}(s)}{\mu_{0,0,0}(s)}, \frac{\mu_{0,0,1}(s)}{\mu_{0,0,0}(s)} \right)$
Distance to Centroid	$\mathcal{D}(s) := \left(\sqrt[2]{\frac{\bar{\mu}_{2,0,0}(s)}{\mu_{0,0,0}(s)}}, \sqrt[2]{\frac{\bar{\mu}_{0,2,0}(s)}{\mu_{0,0,0}(s)}}, \sqrt[2]{\frac{\bar{\mu}_{0,0,2}(s)}{\mu_{0,0,0}(s)}} \right)$

4.2.2.3 Test-time adaptation and inference with shape moments constraints

Given a single new subject $\mathcal{I}_N$ in the target domain composed of slices $I_n : \Omega_t \subset \mathbb{R}^2 \rightarrow \mathbb{R}$, $n = 1, \ldots, N$, and its associated softmax prediction $S(\theta)$, the first loss term in our adaptation phase is derived from Wang *et al.* (2021), to encourage high confidence in the softmax predictions, by minimizing their weighted Shannon entropy:

$$\ell_{ent}(\mathbf{s}_n(i,\theta)) = -\sum_k \nu_k s_n^k(i,\theta) \log s_n^k(i,\theta), \tag{4.5}$$

where $\nu_k, k = 1 \ldots K$, are class weights added to mitigate imbalanced class-ratios.

Ideally, to guide adaptation, we would penalize the deviations between the shape descriptors of the softmax $S(\theta)$ and those corresponding to the ground truth $\mathcal{Y}$. As the ground-truth labels are unavailable, instead, we estimate the shape descriptors from textbook anatomical knowlegde, which we denote respectively $\bar{\mathcal{R}}, \bar{\mathcal{C}}, \bar{\mathcal{D}}$ (see I).

We first constrain the total volume of the shape of each class k, by including the simplest zero-order moment: the class-ratios $\mathcal{R}$. Seeing these class ratios as distributions, we integrate a KL divergence with the Shannon entropy:

$$\mathcal{L}_{TTAS}(\theta) = \sum_n \frac{1}{|\Omega_n|} \sum_{i \in \Omega_t} \ell_{ent}(s_n(i,\theta)) + \mathrm{KL}(\mathcal{R}(S(\theta)), \bar{\mathcal{R}}). \tag{4.6}$$

It is worth noting that, unlike Bateson *et al.* (2022a), which used a loss of the form in Eq (4.6) for training on target data, here we use this term for inference on a test subject, as a part of our overall shape-based objective. Additionally, we embed the centroid ($\mathcal{M} = C$) and/or the distance to centroid ($\mathcal{M} = \mathcal{D}$) to further guide adaptation to plausible solutions:

$$\begin{aligned} \min_{\theta} \quad & \mathcal{L}_{TTAS}(\theta) \\ \text{s.t.} \quad & \left| \mathcal{M}^{(k)}(S(\theta)) - \bar{\mathcal{M}}^{(k)} \right| \leq 0.1, \quad k = \{2, \ldots, K\}. \end{aligned} \tag{4.7}$$

In standard convex-optimization problems, imposing such hard constraints is typically handled through the minimization of the Lagrangian dual, solving primal and dual problems in an alternating scheme Bertsekas (1995). As this is computationally intractable in deep networks, inequality constraints such as Eq (4.7) are typically relaxed to soft penalties He *et al.* (2017); Jia *et al.* (2017); Kervadec *et al.* (2019b). Therefore, we experiment with the integration of one or both of the moments $\mathcal{M}_j \in \{\mathcal{C}, \mathcal{D}\}$ through a quadratic penalty, leading to the following unconstrained objective for joint test-time adaptation and inference:

$$\mathcal{L}(\mathcal{I}_N) = \sum_n \frac{1}{|\Omega_t|} \sum_{i \in \Omega_n} \ell_{ent}(\mathbf{s}_n(i, \theta)) + \mathrm{KL}(\mathcal{R}(S(\theta)), \bar{\mathcal{R}}) + \lambda \sum_j \mathcal{F}(\mathcal{M}_j(S(\theta)), \bar{\mathcal{M}}_j), \quad (4.8)$$

where $\mathcal{F}$ is a quadratic penalty function corresponding to the relaxation of Eq (4.7): $\mathcal{F}(m_1, m_2) = [m_1 - 0.9m_2]_+^2 + [1.1m_2 - m_1]_+^2$ and $[m]_+ = \max(0, m)$. λ denotes a weighting hyper-parameter. Following recent TTA methods Karani *et al.* (2021); Wang *et al.* (2021), our method updates the batch normalization statistics and only optimizes for the scale and bias parameters of batch normalization layers while the rest of the network is frozen. Figure 4.1 shows the overview of the proposed framework.

4.3 Experiments

4.3.1 Datasets

4.3.1.1 Heart Application

We first evaluate on the 2017 Multi-Modality Whole Heart Segmentation (MMWHS) challenge dataset for cardiac segmentation Zhuang *et al.* (2019). The dataset consists of 20 MRI (source domain) and 20 CT volumes (target domain) of non-overlapping subjects, with their manual

annotations of four cardiac structures: the Ascending Aorta (AA), the Left Atrium (LA), the Left Ventricle (LV) and the Myocardium (MYO). Each source volume has a resolution of $0.78 \times 0.78 \times 1.6$ mm/voxel, while each target volume has a resolution of about $1 \times 1 \times 1$ mm/vox. We employ the pre-processed data provided by Dou *et al.* (2019). In order for the volumes to be roughly on the same view, they cropped the original scans to center the structures to segment using a 3D bounding box with a fixed coronal plane size of 256×256. The scans were normalized as zero mean and unit variance, and data augmentation based on affine transformations was performed. For the domain adaptation benchmark methods (DA and SFDA), we use the data split in Dou *et al.* (2019): 14 subjects for training, 2 for validation, and 4 for testing. Each subject has $N = 256$ slices.

4.3.1.2 Prostate Application

We also evaluate on the dataset from the publicly available NCI-ISBI 2013 Challenge[13]. It is composed of manually annotated T2-weighted MRI from two different sites. The source dataset consists of 30 volumes from Radboud University Nijmegen Medical Centre, generated with a 3T Siemens scanner. Each source volume has a resolution of $0.4 \times 0.4 \times 3$ mm/vox. The target dataset consists of 30 volumes from Boston Medical Center acquired with a 1.5T Philips Achieva. Each target volume has a resolution of 0.6-0.625 $\times$ 0.6-0.625 $\times$ 3.6-4 mm/vox. For the DA and SFDA benchmark methods, 19 scans were used for training, one for validation, and 10 scans for testing. We used the pre-processed dataset from Liu *et al.* (2020), who resized each sample to 384×384 in axial plane, and normalized it to zero mean and unit variance. We employed data augmentation based on affine transformations on the source domain. Each subject has $N \in [15, 24]$ slices.

4.3.2 Benchmark Methods

We experiment with the inclusion of one or multiple shape constraints through shapes moments as described in Section 4.2.1. These shape moments can be derived either in 2D, i.e. for each

[13] https://wiki.cancerimagingarchive.net

slice (we denote this class of models $TTAS_2$) or in 3D for the whole 3D volume (denoted $TTAS_3$). Our first model denoted $TTAS_2(\mathcal{R})$ (resp. $TTAS_3(\mathcal{R})$) constrains only the 2D (resp. 3D) class-ratio $\mathcal{R}$. Our model $TTAS_2(\mathcal{RC})$ constrains $\mathcal{R}$ and the centroid C; $TTAS_2(\mathcal{RD})$ constrains $\mathcal{R}$ and the distance-to-centroid $\mathcal{D}$, and $TTAS_2(\mathcal{RCD})$ constrains $\mathcal{R}, C, \mathcal{D}$ in 2D. Similarly, we derive the 3D models $TTAS_3(\mathcal{RC}), TTAS_3(\mathcal{RD}), TTAS_3(\mathcal{RCD})$.

We compare to two test-time adapation (*TTA*) methods: the first is $Tent$ Wang *et al.* (2021), which is based on the following loss: $\min_\theta \sum_n \sum_{i \in \Omega_n} \ell_{ent}(\mathbf{s}_n(i, \theta))$. Note that $Tent$ corresponds to performing an ablation of all shape moments terms in our loss. We also compared to the method in Karani *et al.* (2021), denoted $TTDAE$, where an additional training task in the source training phase is introduced, which consists in an auxiliary network to denoise segmentation masks. We also implemented two *DA* methods based on class-ratio matching, *CDA* Bateson *et al.* (2021), and $CurDA$ Zhang *et al.* (2020a), and to the recent source-free domain adaptation (*SFDA*) method $AdaMI$ in Bateson *et al.* (2022a). Note that in these methods, the class-ratio matching was implemented in 2D, therefore we kept this implementation in our benchmark. A model trained on the source only, *NoAdap*, was used as a lower bound. A model trained on the target domain with the cross-entropy loss, $Oracle$, served as an upper bound.

4.3.3 Test-time Adaptation with shape descriptors

4.3.3.1 Slice-based 2D constraints

In the 2D version of our method, the moments are computed on each slices, and the moments constraints are also slice-based. For the estimation of the 2D class-ratio $\bar{\mathcal{R}}$, we employed a coarse estimation which is derived from anatomical knowledge available in the clinical literature (see I). For $\mathcal{M} \in \{C, \mathcal{D}\}$, we estimate the shape descriptors using the predictions from the whole subject $\{S_n(\theta), n = 1, \ldots, N\}$. Specifically, we estimate the 2D target shape descriptor from the network prediction masks $\hat{\mathbf{y}}_\mathbf{n}$ after each epoch: $\bar{\mathcal{M}}^{(k)} = \frac{1}{|V^k|} \sum_{v \in V^k} v$, with $V^k = \left\{ \mathcal{M}^{(k)}(\hat{\mathbf{y}}_\mathbf{n}) \text{ if } \mathcal{R}^k(\hat{\mathbf{y}}_\mathbf{n}) > \epsilon^k, n = 1 \cdots N \right\}$.

In this 2D version of our method, we additionally use weak supervision. This takes the form of simple image-level tags by setting $\bar{\mathcal{R}}^{(k)}(\hat{\mathbf{y}}_\mathbf{n}) = \mathbf{0}$ and $\lambda = 0$ if the target image I_n does not contain structure k. Note that the exact same estimation for the class-ratio priors and weak supervision were employed in the benchmarks methods in Bateson *et al.* (2021,2); Zhang *et al.* (2020a).

4.3.3.2 Global 3D constraints

In the 3D version of our method, the moments are computed on each 3D volume, and the moments constraints are also volume-based. Importantly, this relieves from the need for 2D image-level tags such as in Section 4.3.3.1. For the estimation of the 3D class-ratio $\bar{\mathcal{R}}$ and the distance to centroid $\bar{\mathcal{D}}$, we employed an estimation which was derived from textbook anatomical knowledge (see I). Regarding the centroid $\bar{C}$, obtaining its position from such prior knowledge is often impossible. Indeed, although in some cases, we can assume that the organ of interest will be centered in the 3D volume acquired, this is not the case in multi-class applications, and when the 3D volume undergoes data augmentation steps such as translation and rotation. Thus, the centroid position of the anatomical structures in a 3D volume is dependent on the acquisition and the preprocessing steps. Therefore, instead of trying to estimate it, we study the scenario where $\bar{C}$ is obtained through a simple user input of the centroid, with an imprecision of $\pm 10\%$ for each 3D coordinate.

4.3.3.2.1 Stochastic estimation of the gradient:

Constraining the shape of the whole 3D segmentation mask in the network loss requires to compute the loss function $\mathcal{L}(\mathcal{I}_N)$ in Eq (4.8) and backpropagate its gradient $\mathcal{G}(\mathcal{I}_N)$ on the whole 3D volume. In practice, this is memory intensive, and will not feasible with most GPUs commonly available. Instead, we use a stochastic estimation of $\mathcal{G}(\mathcal{I}_N)$, which we detail in Algorithm 4.1.

Algorithm 4.1 Test-Time Adaptation with 3D shape constraints

Initialization: *The affine transformation parameters* $\{\gamma_{l,c}, \beta_{l,c}\}$ *for each normalization layer* l *and channel* c *in the source model are collected. The remaining parameters* $\theta \backslash \{\gamma_{l,c}, \beta_{l,c}\}$ *are frozen. The normalization statistics* $\{m_{l,c}, \sigma_{l,c}\}$ *from the source data are discarded.*

Require: *A single new 3D subject composed of N slices* $\mathcal{I}_N = \{I_n : \Omega_t \subset \mathbb{R}^2 \rightarrow \mathbb{R}, n = 1, \ldots, N\}$. **Require:** $Q \geq 0$ *the chosen active batch size.*

Require: $\mathcal{L}$ *the loss function to optimize. Denote* $\mathcal{G}$ *its gradient.*

for $j \geq 0$ **do**

  while $(j+1) \times Q \leq N$ **do**

    update active batch A_j and frozen batch F_j as following:

    $A_j \leftarrow I_n, n = j * Q, \ldots, (j+1)Q$ has active gradients.

    $F_j \leftarrow \mathcal{I}_N \backslash A_j$ has frozen gradients.

    Forward pass of each batch A_j and F_j to obtain the network prediction on the whole 3D volume Compute $\mathcal{L}(\mathcal{I}_N)$ ▷ Eq. 4.8

    Compute $\mathcal{G}(A_j)$

    Backpropagation using $\mathcal{G}(A_j)$ to update $\{\gamma_{l,c}, \beta_{l,c}\}$.

  end

end

Termination: *Note that the transformation update follows the prediction for the current batch. Therefore the model is first updated and then inference is repeated. This adaptation process is repeated for multiple epochs*

Table 4.2 Test-time metrics on the cardiac dataset, for our method and various *Domain Adaptation* (DA), *Source Free Domain Adaptation* (SFDA) and *Test Time Adaptation* (TTA) methods

Methods	DA	SFDA	TTA	Target Supervision	DSC (%)					ASD (vox)				
					AA	LA	LV	Myo	Mean	AA	LA	LV	Myo	Mean
NoAdap (lower b.)				×	49.8	62.0	21.1	22.1	38.8	19.8	13.0	13.3	12.4	14.6
Oracle (upper b.)				Full Supervision	91.9	88.3	91.0	85.8	89.2	3.1	3.4	3.6	2.2	3.0
CurDA Zhang *et al.* (2020a)	✓	×	×	2D Tags	79.0	77.9	64.4	61.3	70.7	6.5	7.6	7.2	9.1	7.6
CDA Bateson *et al.* (2021)	✓	×	×	2D Tags	77.3	72.8	73.7	61.9	71.4	4.1	6.3	6.6	6.6	5.9
AdaMI Bateson *et al.* (2022a)	×	✓	×	2D Tags	83.1	78.2	74.5	66.8	75.7	5.6	**4.2**	**5.7**	6.9	5.6
TTDAE Karani *et al.* (2021)	×	×	✓	×	59.8	26.4	32.3	44.4	40.7	15.1	11.7	13.6	11.3	12.9
Tent Wang *et al.* (2021)	×	×	✓	×	55.4	33.4	63.0	41.1	48.2	18.0	8.7	8.1	10.1	11.2
Proposed Method with 2D Shape Moments														
TTAS$_2(\mathcal{R})$	×	×	✓	2D Tags	78.9	77.7	74.8	65.3	74.2	5.2	4.9	7.0	7.6	6.2
TTAS$_2(\mathcal{RC})$	×	×	✓	2D Tags	**85.1**	**82.6**	79.3	73.2	**80.0**	5.6	**4.3**	6.1	5.3	5.3
TTAS$_2(\mathcal{RD})$	×	×	✓	2D Tags	82.3	78.9	76.1	68.4	76.5	**4.0**	5.8	6.1	5.7	5.4
TTAS$_2(\mathcal{RCD})$	×	×	✓	2D Tags	84.9	80.3	**79.7**	**73.8**	79.7	4.2	4.8	**5.9**	**5.2**	**5.0**
Proposed Method with 3D Shape Moments														
TTAS$_3(\mathcal{R})$	×	×	✓	×	66.5	67.4	63.0	62.9	65.0	7.6	7.4	8.3	8.6	8.0
TTAS$_3(\mathcal{RC})$	×	×	✓	3D centroid estimation	82.8	73.9	70.4	55.7	70.7	4.9	6.5	6.6	8.4	6.6
TTAS$_3(\mathcal{RD})$	×	×	✓	×	87.3	76.5	67.0	57.2	72.0	5.0	4.4	7.7	8.1	6.3
TTAS$_3(\mathcal{RCD})$	×	×	✓	3D centroid estimation	**80.6**	**74.2**	**68.0**	**71.0**	**73.4**	4.9	3.9	6.2	8.2	5.8

Table 4.3 Test-time metrics on the prostate dataset

Methods	DA	SFDA	TTA	Target Supervision	DSC (%)	ASD (vox)
NoAdap (lower bound)				×	67.2	10.60
Oracle (upper bound)				Full Supervision	88.9	1.88
CurDA Zhang *et al.* (2020a)	✓	×	×	2D Tags	76.3	3.93
CDA Bateson *et al.* (2021)	✓	×	×	2D Tags	77.9	3.28
AdaMIBateson *et al.* (2022a)	×	✓	×	2D Tags	79.5	3.92
TTDAE Karani *et al.* (2021)	×	×	✓	×	73.2	5.80
Tent Wang *et al.* (2021)	×	×	✓	×	68.7	5.87
Proposed Method with 2D Shape Moments						
$TTAS_2(\mathcal{R})$	×	×	✓	2D Tags	75.3	5.06
$TTAS_2(\mathcal{RC})$	×	×	✓	2D Tags	80.2	3.79
$TTAS_2(\mathcal{RD})$	×	×	✓	2D Tags	79.5	3.90
$TTAS_2(\mathcal{RCD})$	×	×	✓	2D Tags	**81.2**	3.69
Proposed Method with 3D Shape Moments						
$TTAS_3(\mathcal{R})$	×	×	✓	×	70.5	5.41
$TTAS_3(\mathcal{RC})$	×	×	✓	3D centroid estimation	74.3	4.85
$TTAS_3(\mathcal{RD})$	×	×	✓	×	78.2	3.70
$TTAS_3(\mathcal{RCD})$	×	×	✓	3D centroid estimation	78.5	**3.57**

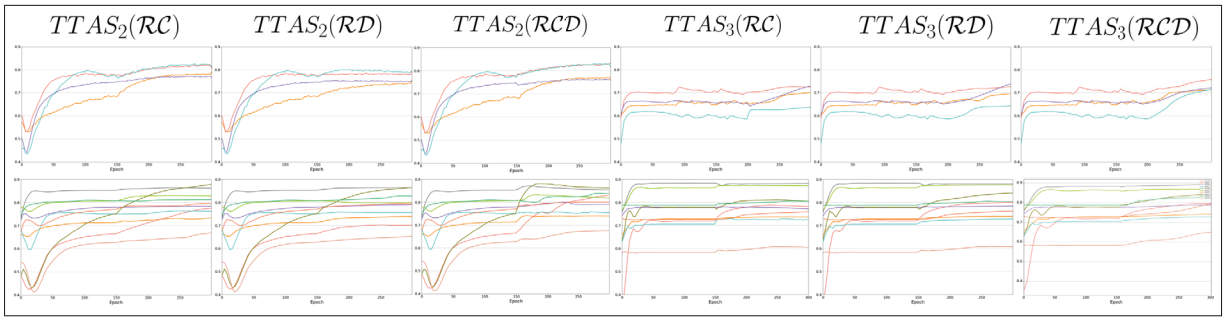


Figure 4.2 Per-subject DSC performance throughout epochs on cardiac images (top) and prostate images (bottom), shown for $TTAS_2$ (left) and $TTAS_3$ (right), which constrain the 2D (resp. 3D) class-ratio $\mathcal{R}$, the centroid $\mathcal{C}$ and/or the distance-to-centroid $\mathcal{D}$ of structures. Each color represents the performance for a single subject. Our method subsequently improves the DSC in both applications, for each suject.

4.3.3.3 Training and implementation details

For all methods, we employed 2D U-Net as the segmentation network Ronneberger *et al.* (2015). Note that a 3D U-Net could have been used for the $TTAS_3$ models, however 3D U-Nets are known for their memory issues and 2D U-Net commonly yield state-of-the-art segmentation performance. A model trained on the source data with Eq (4.4) for 150 epochs was used

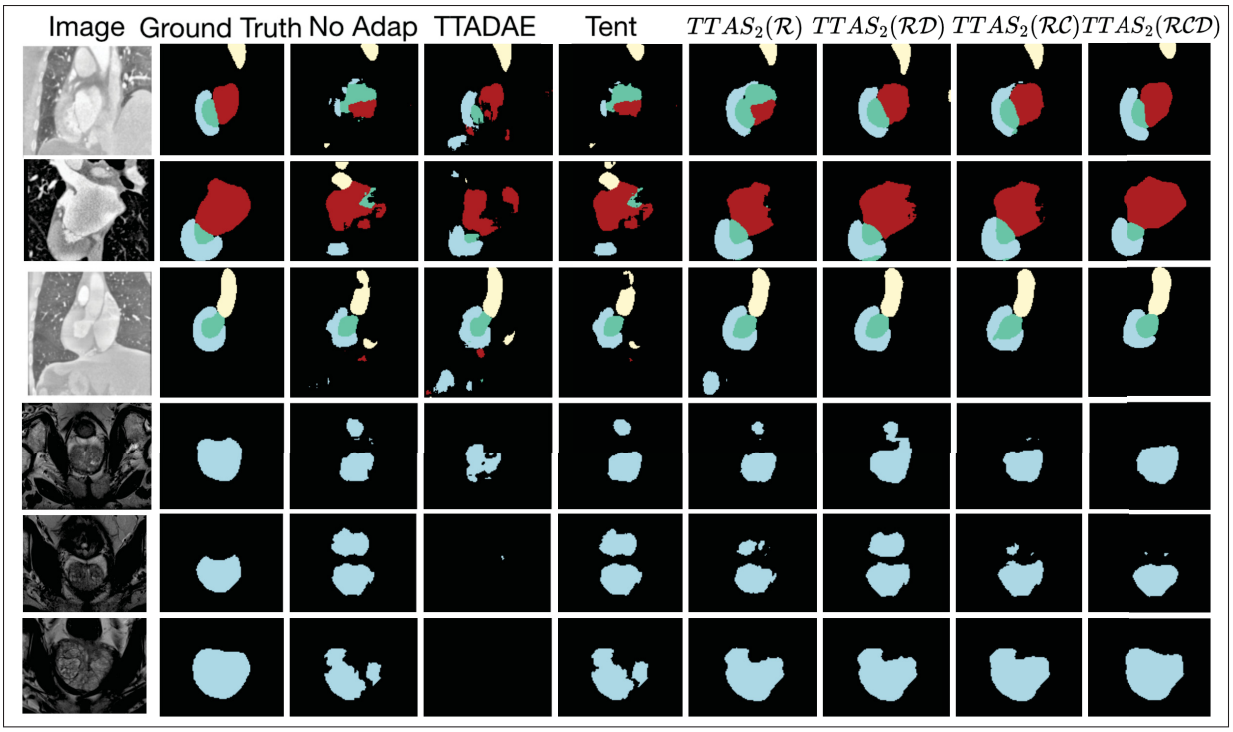


Figure 4.3 Qualitative performance on cardiac images (top) and prostate images (bottom): examples of the segmentations achieved by our formulations ($TTAS_2(\mathcal{RC}), TTAS_2(\mathcal{RD}), TTAS_2(\mathcal{RCD})$), and benchmark TTA models. The cardiac structures of MYO, LA, LV and AA are depicted in blue, red, green and yellow respectively. Visual results confirm the substantial effect of shape-moment priors on TTA performances.

as initialization. For DA and SFDA models, the model is retrained on a set *Tr* of target training samples, and model validation is done on a target set *Tv*. On the contrary, test-time adaptation (TTA) models do not access *Tr* nor *Tv*, and adaptation is performed on each test subject independently. For all TTA methods, we estimate the normalization statistics during testing on the target data. Our TTAS model was initialized with Eq (4.6) for 150 epochs, after which the additional shape constraints were added using Eq (4.8) for 150 epochs. As *Tr* and *Tv* are inaccessible, the learning parameters are set following those in the source training, and the hyper-parameters are fixed across experiments: we trained with the Adam optimizer Kingma & Ba (2014), an initial learning rate of 5×10^{-4}, a learning rate decay of 0.9 every 20 epochs, and a weight decay of 1×10^{-4}. The weights ν_k are calculated as: $\nu_k = \frac{\bar{\mathcal{R}}_k^{-1}}{\sum_k \bar{\mathcal{R}}_k^{-1}}$. We set $\lambda = 1 \times 10^{-4}$ for all experiments. For all methods, we pick the final model as the one at

the last epoch. For the $TTAS_2$ models with slice-based 2D constraints, we set the batch size to *min(N,22)*, and ϵ to 1×10^{-2}. For the $TTAS_3$ models with global 3D constraints, we set the active batch size $Q = N/4$ (see Algorithm 4.1).

4.3.3.4 Evaluation

We use the 3D Dice similarity coefficient (DSC) and the 3D Average Surface distance (ASD) as evaluation metrics in our experiments.

4.3.4 Quantitative Results

Table 4.2 and Table 4.3 report quantitative metrics for the heart and prostate applications respectively. As expected, we see that the models trained with full supervision on the source domain suffer from a drop in performance when used in a different target domain without any adaptation.

4.3.4.1 Comparison to DA and SFDA methods

All models that use domain adaptation (DA), or source-free domain adaptation (SFDA) yield a substantial improvement over the lower baseline. The source-free $AdaMI$ achieves the best DSC improvement over the lower baseline *NoAdap*, with a mean DSC of 75.7% (cardiac) and 79.5% (prostate). It is interesting to compare $AdaMI$ to $TTAS_2(\mathcal{R})$, as they optimize the same loss function, but on different subjects. $AdaMI$ only slightly outperforms $TTAS_2(\mathcal{R})$ on the cardiac application, while the improve is greater for prostate one. However, adding an additional shape moment such as the distance-to-centroid in $TTAS_2(\mathcal{RD})$ yields better scores then $AdaMI$: 76.5% DSC, 5.4 vox. ASD (cardiac) and 79.5% DSC, 3.9 vox. ASD (prostate); while $TTAS_2(\mathcal{RC})$ achieves an even better performance: 80.0% DSC and 5.3 vox. ASD (cardiac), 80.2% DSC and 3.79 ASD vox. (prostate). Moreover, the efficacy of adding multiple shape moments is confirmed by the model $TTAS_2(\mathcal{RCD})$, with 79.5% DSC (cardiac) and 81.9% DSC (prostate), and the best ASD out of all 2D methods: 5.0 vox. ASD (cardiac), 3.69 ASD vox

(prostate). This is remarkable, as all these models have access to the same annotations, i.e. 2D image-level tags, and use the same class-ratio prior. This points to the efficiency of TTA framework for adapting to new target subjects, compared to the more common DA and SFDA framework.

4.3.4.2 Comparison to TTA methods

On both applications, $TTAS$ outperforms the two other single-subject test-time domain adaptation methods. This is the case for both the 2D and the 3D version of our method. Specifically, *TTDAE* yields a smaller improvement than $TTAS$ on the prostate applications, with 73.2% DSC, 5.80 vox. ASD. On the more difficult heart one, TTDAE fails to perform adaptation, only reaching 40.7% DSC, 12.9 vox. ASD. Regarding *Tent*, a small improvement on both applications is obtained: 48.2% DSC, 11.2 vox. ASD (cardiac) and 68.7% DSC, 5.87 vox. ASD (prostate). This shows that combining shape moments with entropy minimization as in our $TTAS_2$ yields a stronger test-time adaptation model.

4.3.4.3 Comparing 2D versus 3D constraints

Comparing the 2D and 3D versions of our method ($TTAS_2$ and $TTAS_3$), we see that on both applications, $TTAS_2$ outperforms $TTAS_3$. However, the annotations that both model have access to should also be compared: $TTAS_2(\mathcal{RD})$ uses slice-level tags informing on the presence or absence of structures. Instead $TTAS_3(\mathcal{R})$, $TTAS_3(\mathcal{RD})$ are fully unsupervised, and $TTAS_3(\mathcal{RC})$, $TTAS_3(\mathcal{RCD})$ use an estimation of the 3D centroid of structures. Thus, $TTAS_3$ requires much less annotations, while achieving a close performance in all experiments. For instance, the best 3D model $TTAS_3(\mathcal{RCD})$, yields a 6.3% DSC drop for cardiac (resp. 2.9% for prostate) only compared to the best 2D model. This highlights the effectiveness of our extension to 3D constraints, $TTAS_3$, compared to our previous $TTAS_2$ Bateson, Lombaert & Ayed (2022b).

Finally, the learning curves in Fig. 4.2 illustrate the stability of the optimization process for $TTAS$: specifically, the per-subject DSC scores are shown throughout epochs for the prostate

application (top) and the cardiac one (bottom), both for $TTAS_2$ (left) and $TTAS_3$ (right). First, we can observe that on both applications, the DSC improvement is substantial for each individual subject without exception. Moreover, after a few epochs, the performance of $TTAS$ improves steadily before reaching convergence for each subject, and each combination of shape moments. This convergence property is particularly important as there is no validation set, and therefore no stopping criteria for the optimization procedure. This suggests that the shape moments matching helps stabilizing the adaptation process.

4.3.5 Qualitative Results

Qualitative segmentations are depicted in Figure 4.3. These visuals results confirm that without adaptation, a model trained on source data only cannot properly segment the structures on the target images. The segmentation masks obtained using the TTA formulations $Tent$ Wang *et al.* (2021), $TTADAE$ Karani *et al.* (2021) only show little improvement. Both methods are unable to recover existing structures when the initialization $NoAdap$ fails to detect them (see fourth and fifth row, Figure 4.3). On the contrary, those produced from our model constraining the class-ratio only, $TTAS_2(\mathcal{R})$ show more regular edges and is closer to the ground truth. However, the improvement over $TTAS_2(\mathcal{R})$ obtained by our models $TTAS_2(\mathcal{RC})$, $TTAS_2(\mathcal{RD})$, is remarkable regarding the shape and position of each structures: the prediction masks show better centroid position (first row, see LA and LV) and better compactness (third, fourth, fifth row, Figure 4.3). Finally, we visually confirm the usefulness of integrating multiple shape-moment priors such as in $TTAS_2(\mathcal{RCD})$. Indeed, this model able to recover the structures with less aberrations, and better fitted shapes in both applications (second, fourth, fifth row).

4.4 Discussion

We have introduced a shape-driven test-time domain adaptation (TTAS) method to guide a segmentation network, trained on a source domain, to perform on a single subject from a different target domain. In our setting, the model does not have any access to the source-domain data or to target samples for retraining in the adaptation phase. We have demonstrated the robustness of

our TTAS approach on cross-site prostate MRI, and MRI to-CT cardiac adaptation. Indeed, the segmentation performance has been consistently and significantly improved in both tasks, for each individual subject in the target modality.

Test-Time Adaptation: Surprisingly, our model yields comparable or better performance than state-of-the art domain adaptation approaches that do rely on both source and target data for re-training (Bateson *et al.* (2021), Zhang *et al.* (2020a)). We have also demonstrated the superiority of our method when compared to the source-free domain adaptation method *AdaMI* Bateson *et al.* (2022a), which adapts by regularizing the zero-order shape moment, i.e. the class-ratio on a dedicated set of target samples. These results suggest that having access to source samples and to target training samples may not be necessary for adaptation. Instead adapting directly to a subject from the target domain can be more efficient than adaptation methods which have seen more data of the target domain but not this specific subject.

Shape moments constraints: Our method $TTAS$ adapts a given segmentation network to a new subject from a different target domain via shape regularization. We embed such regularization by constraining the shape moments of the predicted segmentation masks. In particular, we have shown the efficiency of constraints on the size area, the centroid, and the distance-to-centroid of anatomical structures. We have also confirmed the usefulness of integrating all of these shape moments together such as in $TTAS(\mathcal{RCD})$. In our experiments, it is observed that imposing constraints of increasing order (up to 2nd order) enables to better restore the main characteristics of the shape (size, position, covariance structure...). This is even more remarkable for the cardiac application, for which the substructures exhibit significant variations in their size and shape. Finally, we highlight that in contrast to many works on shape priors in segmentation, the use of shape moments does not require shape learning. Therefore, it could be easily applied to arbitrary shapes since these moment constraints do not affect the recovery of fine scale details.

Embedding $TTAS$ in 2D and 3D: We have implemented our formulation by embedding the shape-moment constraints either in 2D (slice-based), or in 3D (volume-based). Our experiments

confirm the efficiency of both implementations. Notably, our method is able to perform test-time adaptation on a whole 3D volume, but does not require the initial network to be trained in 3D.

Robustness and stability: Our experiments have further shown that *TTAS* is robust to a substantial imprecision of the shape-moment priors. In our implementation, the shape moments are derived from readily available anatomical reference values. Therefore, having a coarse knowledge of the target shape moments seems to be enough to guide adaptation. Moreover, regarding convergence and stability properties, these are well-known challenges for unsupervised and weakly supervised test-time adaptation methods. We address them by only updating the batch normalisation statistics and parameters of the segmentation network. We have verified that this ensures the stability and convergence of our model adaptation *on each new subject* of the target domain, in both applications. Importantly, the general efficacy of our method has been demonstrated without per-subject nor per-application tuning of learning parameters and hyperparameters.

Integration of anatomical knowledge and annotations: Importantly, our approach is able to flexibly integrate different levels of annotations in the target domain, from no supervision to weak supervision (e.g. image-level tags, 3D centroid) that informs on the shape moments. We believe this can have great practical value in many medical applications. Indeed, such weak supervision can be commonly accessible in practice, thanks to routine clinical annotations, conventions in patient position, and/or anatomical stability. Therefore, integrating shape moments such as in our framework allows to make use of a wide range of anatomical knowledge and annotations, with different levels of precision. In the applications we tackle, the shape moments could be further precised by common annotations, such as subjects characteristics (gender, age), or the phase of the cycle (diastole or systole) in which the cardiac image was acquired.

Extension to higher order moments: In our experiments on both applications, we also tested the integration of higher-order shape moments. However, we have found that they do not yield significant improvement compared to the zero, first and second order moments that we integrate here. This is in line with the work of Klodt & Cremers (2011a) on integrating shape moments

in non-deep convex models, which found that the shape improvements due to higher order constraints are fairly small. They further note that imposing higher order moments is impractical, as the user cannot estimate these moments visually, nor are they easy to derive from textbook anatomical knowledge or from common clinical annotations.

Limitations: One limitation of our work is that constraining the predicted shape of anatomical structures via moments may not be relevant in some applications where these shapes are extremely variable, and/or when an estimation of these moments via anatomical knowledge or annotations is impossible. Such an example could be lesion segmentation tasks Kamnitsas *et al.* (2017), where the size, shape, and position of lesions often cannot be known in advance without clinical annotations.

Future work: Amongst interesting extensions left for future work could be the validation of our method in challenging TTA settings with known drastic population shifts in the test set, e.g. differences in age, gender, clinical diagnosis. We argue that such shifts could be well taken into account by the specification of one or various of the shape moments that we integrate here (size, centroid, distance-to-centroid). An example for the cardiac application could involve training on healthy subjects (source domain) and testing on subjects with hypertrophic cardiomyopathy, which is known to enlarge the left ventricule (LV), thereby modify its low-order shape moments. Similarly, for the prostate application, an example could involve testing on subjects from a different age group or with clinical diagnosis (e.g. prostate enlargment), as the size of the prostate (the zero-order moment) is the main biomarker in both cases.

4.5 Conclusion

In this paper, we proposed a simple formulation for *single-subject* test-time adaptation (TTA), which does not need access to the source data, nor the availability of a set of target training samples. Our method does not require any modification of the source training stage, contrary to the recent TTA method Karani *et al.* (2021). Instead, it operates from an off-the-shelf source pre-trained segmentation model, using shape regularization on a single 3D data point in the

target domain. Specifically, our approach minimizes the entropy of predictions and a class-ratio prior over batch normalization parameters. To further guide adaptation, we integrate shape moments through penalty constraints. We obtain an estimation of these moments from available textbook anatomical knowledge. We show that multiple shape moments can be successfully integrated to achieve more reliable segmentations predictions, with less aberrations and more plausible shapes. We validate our method on two challenging tasks, the MRI-to-CT adaptation of cardiac segmentation and the cross-site adaptation of prostate segmentation. Our formulation achieved better performances than state-of-the-art TTA methods, with a 31.8% (resp. 8.2%) DSC improvement on cardiac and prostate images respectively. Surprisingly, it also fares better than various state-of-the-art domain adaptation methods. These results highlight the effectiveness of shape moments on test-time inference, and question the usefulness of training on target data in segmentation adaptation. Our test-time adaptation framework is straightforward to use with any segmentation network architecture.

CONCLUSION AND RECOMMENDATIONS

The literature reviewed in the background chapter highlights the challenges of domain adaptation for image segmentation and the unrealistic setting of most state-of-the-art domain adaptation frameworks. This thesis addresses these issues by proposing a set of methods to adapt deep segmentation networks with less data. Specifically, the three research objectives led to novel tools useful for integrating prior domain knowledge into the adaptation framework of deep segmentation models to compensate for the missing data. In this last chapter of the thesis, each contribution for the three objectives is summarized with its practical impact. We then address current limitations of our contributions and possible directions for future works.

5.1 Summary of contributions

5.1.1 Objective 1: Constrained domain adaptation for image segmentation

In chapter 2, we have proposed a general novel methodology for adapting deep segmentation models based on a constrained formulation that incorporates prior knowledge of the target segmentation regions. Such knowledge may take the form of anatomical information, such as structural shape. In our experiments, we show the efficiency of a simple size prior. Our general approach imposes inequality constraints on the network predictions of unlabeled or weakly labeled target samples, with acceptable prior knowledge uncertainty. We handle the ensuing constrained optimization problem with differentiable penalties, suited for stochastic gradient descent approaches.

Impact: Compared to state-of-the-art adversarial methods for adapting deep segmentation models, our method greatly reduces computationally complexity, by optimizing a single network, without reducing adaptation performance. Moreover, our formulation opens up new avenues to incorporate a wide variety of anatomical constraints. The findings in this chapter have the potential to improve the robustness to domain shift of segmentation networks with applications in numerous medical applications.

5.1.2 Objective 2: Source-free domain adaptation for image segmentation

In chapter 3, we introduced a source-free domain adaptation method, where the source images are entirely absent in the adaptation phase, doing away with the unrealistic data availability setting of most DA frameworks. By studying how to best leverage anatomical shape knowledge, our formulation minimized a label-free entropy loss defined over target-domain data, which we further guide with a domain-invariant size prior on the segmentation regions. Importantly, our experiments on prostate, heart, and spine images reflect the broad applicability of our method with widely different domain shifts and anatomical structures.

Impact: The proposed source-free domain method overcomes a major limitation of current adaptation approaches, i.e. the requirement for concurrent access to source and target data. Our method has already proved to be a useful contribution for the community, and is well-recognized as one of the first source-free domain adaptation frameworks for image segmentation. It has received many positive feedbacks and reports of improved performances on various tasks Chen, Liu, Jin, Dou & Heng (2021a); Kundu *et al.* (2021); Liu, Xing, Yang, El Fakhri & Woo (2021c).

5.1.3 Objective 3: Single-subject test-time adaptation with shape moments for segmentation

In chapter 4, we push further the complexity of the adaptation task and its robustness to missing data, introducing a single-subject test-time adaptation framework. Our method relies on high-level shape knowledge of the target structures. We have demonstrated the efficiency of integrating various shape moments in combination with entropy minimization. We have validated our method with both local constraints over 2D slices and global constraints over 3D image domains. The quantitative performance of our single-subject adaptation method is on-par or better than state-of-the-art methods in the literature which access *whole data distributions* both in the source and in the target domain, and are much more complex. Moreover, the introduction of high-level shape moments leads to more plausible predicted shapes.

Impact: In clinical practice, the single-subject test-time adaptation framework of chapter 4 is by far the most plausible among common DA adaptation settings. Test-time adaptation is gaining traction in the medical imaging community, and our work is widely acknowledged as one of the earliest test-time adaptation frameworks for image segmentation. The integration of multiple high-level shape moments is promising to improve the robustness of models in many applications, and could also lead to a better explainability of the deep segmentation models.

5.2 Recommendations

The contributions of this thesis are discussed in the previous section. However, some limitations remain that were not thoroughly investigated. In this section, the main shortcomings are identified, and recommendations for future works are provided.

5.2.1 Integrating other priors into the DA framework

In this thesis, the integration of domain knowledge information has been restrained to incorporating low and high order shape moments in the network optimization loss. Yet domain kwowledge could be included in the DA optimization loss through a wide variety of shape and appearance priors which could be further explored. We highlight a few of these priors below.

5.2.1.1 Topological, region interactions, and other shape priors

Anatomical structures in medical imaging typically follow a specific topology that must be preserved to obtain plausible results. Thus, when segmenting multiple anatomical objects, geometric relationships such as containment, exclusion, adjacency should be retained Nosrati & Hamarneh (2016). Similarly, high-level shape priors such as the star-shape Mirikharaji & Hamarneh (2018), tightness priors Kervadec *et al.* (2022) and k-convex priors Isack, Gorelick, Ng, Veksler & Boykov (2018) are efficient descriptive tools in many segmentation applications. The few existing works incorporating such priors to deep network losses are limited to in-domain segmentation tasks

Cui, Wang, Lin, Zhou, Eberl, Feng et al. (2016); Gupta, Hu, Kaan, Jin, Mpoy, Chung et al. (2022); Reddy, Gopinath & Lombaert (2019).

Incorporating these constraints could further improve adaptation models. However, the trade-off between fidelity and optimizability remains problematic when combining numerous and richer shape priors into the adaption loss. Specifically, better modelling the segmentation task usually comes at the expense of optimization simplicity, and often requires heavy hyper-parameters tuning. We advocate for the introduction of more advanced optimization schemes, such as log-barrier methods Kervadec *et al.* (2022) and augmented Lagrangian methods Bertsekas (1976); Sangalli *et al.* (2021) as possible tools to successfully incorporate multiple and rich priors.

5.2.1.2 Atlas as shape priors

Alternatively, more complex shape modelling such as atlas-based priors could be explored Lorenzo-Valdés, Sanchez-Ortiz, Mohiaddin & Rueckert (2002). Although they are not as easily accessible and measurable as high-order shape moments, they can hold rich information regarding the shape of anatomical structures El Jurdi *et al.* (2021); Nosrati & Hamarneh (2016).

Various studies have introduced a probabilistic atlas in supervised organ segmentation tasks Ding, Han & Niethammer (2020); Vakalopoulou, Chassagnon, Bus, Marini Silva, Zacharaki, Revel et al. (2018); Zeng, Karimi, Pang, Mohammed, Schneider, Honarvar et al. (2019), via late fusion strategies, which are sub-optimal. In a recently proposed semi-supervised framework, Huang, Zheng, Lin, Cai, Hu, Zhang et al. (2021) extracted the probabilistic atlas as a prior to guide learning by incorporating the atlas into the loss function. Future work could address the integration of these complex shape priors in the loss functions of unsupervised DA models.

5.2.1.3 Appearance priors

A limitation of our shape-aware adaptation models is that they are not well suited to clinical settings where the shapes to be segmented are extremely variable, and/or when obtaining an

estimation of our high-level priors via anatomical knowledge is impossible. Such an example could be lesion segmentation tasks, where the size, shape, and position of lesions cannot be estimated a priori without clinical annotations. Future work could therefore explore the integration of appearance-based priors such as pixel-intensity and texture, instead of shape priors. As tumors exhibit specific appearance and texture profiles, different from healthy tissue Davnall, Yip, Ljungqvist, Selmi, Ng, Sanghera et al. (2012); Soni, Priya & Bathla (2019), constraining deep networks with appearance priors could help lesion segmentation Gordillo, Montseny & Sobrevilla (2013); Tong, Zhao, Zhang, Chen & Jiang (2019). However, it is not immediately clear how to integrate appearance priors via practical 'texture features'. Statistical approaches, fractal models, or transform-based techniques like Fourier, Gabor, or wavelet transforms are examples of common texture analysis techniques that might be used Nosrati & Hamarneh (2016).

5.2.1.4 Uncertainty and errors in the data, priors, and DA model

Amongst topics left for future exploration is the integration of data and model uncertainty, a challenging task that could draw from recent methodological advances Dusenberry, Tran, Choi, Kemp, Nixon, Jerfel et al. (2020). Data uncertainty refers to incomplete or wrong information and leads to uncertainty in the predicted outcome. The following questions could be studied: what is the impact of imprecision in the domain prior knowledge, such as the shape priors in our applications ? How can the level of required precision on the priors be estimated a priori ? When does an imprecise shape or appearance prior become detrimental to the adaptation task ? Furthermore, in our contributions, the ground truth shape priors are estimated and integrated in a deterministic fashion. Could the imprecision in the domain-knowledge priors be integrated in the framework via a probabilistic approach instead? Regarding model uncertainty, how does the predicted outcome uncertainty differ across patient subgroups? Quantifying uncertainties in the data and model could help clinicians in explaining and interpreting the adaptation model, as well as assessing risk-case situations where domain adaptation could fail.

5.2.2 Pushing further constrained domain adaptation for challenging tasks

5.2.2.1 Longitudinal studies with domain shifts

A contribution of this thesis is to propose an adaptation framework for segmentation networks. While the potential of our method is demonstrated for disjointed subject populations, longitudinal studies with domain shifts could be another relevant application, for instance, a clinical study following patients where the scanner model or acquisition parameters could change in the course of the study. Similarly, the longitudinal study of fetal organ development requires its accurate segmentation at each time step. This challenging task could be handled by our adaptation framework, by specifying various higher-order shape moments (or other shape priors) at each time step. Specifically, as fetal development typically follows a predictable course, we hypothesize that the shape priors also follow predictable changes, which could guide the segmentation task.

5.2.2.2 Population shifts between source and target domain

In this dissertation, we have restrained our applications to domain shifts caused by changes in imaging acquisition (different sites, machines, or modalities). An interesting new direction could be the validation of our method for population shifts. Examples of shifts could be differences in age, gender, and clinical diagnosis between the source and the target domain. In standard deep models, it is not easy to integrate these clinical variables into the prediction framework. For instance, how can a cardiac segmentation network trained on healthy subjects (source domain) be adapted to a target domain comprised only of subjects with hypertrophic cardiomyopathy (HC) and no pixel annotations ? This is not a borderline scenario as it is common clinical practice to collect datasets of symptomatic patients only. Our DA methods aim to integrate domain knowledge in order to adapt a network for each specific target domain. In the example above, HC causes left ventricle enlargement, thereby modifying its low-order shape moments. Future work could therefore explore whether population shifts can be well taken into account by the specification of shape priors such as the shape moments that we have studied.

In the settings described above, the global population shift between the source and the target domain is known. An open question that remains involves dealing with *unknown* population shifts.

5.2.2.3 Causality for domain adaptation

Since deep causal learning discovers underlying causal relationships, and avoids potentially spurious correlations, it holds significant promises to improve the robustness of deep models. A recent line of work aims to mitigate the most significant issues for clinical translations of medical image analysis through causal inference Vlontzos *et al.* (2022). For instance, Schrouff, Harris, Koyejo, Alabdulmohsin, Schnider, Opsahl-Ong et al. (2022) systematically reviews the biases that may occur in the cross-hospital deployment of predictive models in dermatology, suggesting causal analysis as a viable remedy. In this setting, DA becomes *causal* DA, where domain shifts are expected on specific paths of a causal graph. An emerging causal-aware method models the image generating process and includes factors leading to shifts and biases. DA can then be interpreted as a model able to perform under differences in the imaging domain parameter. Once the causal features are modeled and learned, synthetic datasets are created in order to enforce the robustness of methods to interventions on these factors Ouyang, Chen, Li, Li, Qin, Bai et al. (2022).

An ambitious future direction would be to model by causality-aware techniques the common medical imaging domain shifts that we have outlined in Section 1.4.1 (covariate, conditional, label, and annotation shifts). Then, combining them with our DA methods could steer the network towards shape information that is domain-invariant, thereby alleviating the domain shifts effects.

5.2.2.4 Multimodal learning with domain shifts

Medical data is inherently multimodal, with valuable information being routinely acquired and generated, e.g. clinical and radiological reports, genomic sequences, and lab test results.

Recently, a number of studies have demonstrated outstanding results in the joint processing of medical images and clinical reports Heiliger, Sekuboyina, Menze, Egger & Kleesiek (2022). Besides, to emulate the gold standard in cancer survival prediction, multimodal models of histology and genomics are being developed Chen *et al.* (2021b). Methodologies to integrate multimodal data can be categorized into modality fusion (early, joint, and late fusion, depending on the stage in the deep model where the features are combined), representation learning, or modality translation Heiliger *et al.* (2022); Moon, Lee, Shin, Kim & Choi (2022). To the best of our knowledge, no studies have specifically addressed the presence of domain shifts and missing data in such multimodal settings, whether in one or all of the modalities. Owing to the high dimensionality and heterogeneity, these domain shifts are poised to cause significant challenges. Future work could address how to integrate domain knowledge to improve the robustness of deep models in such challenging settings.

In summary, the findings of the thesis provide tools for adapting deep segmentation models to new target domains with missing data. The first research objective led to a general methodology based on constrained domain adaptation, capable of integrating knowledge on the anatomical regions to guide adaptation in the target domain. The second research objective led to the development of a source-free domain adaptation method for deep segmentation networks. The proposed approach was found to be efficient to adapt prostate, cardiac, and spine segmentation networks, ensuring generalizability in a wide range of applications. The third objective enables single-subject test-time adaptation for segmentation networks via shape moments matching. The works proposed in the thesis, together with concrete recommendations for future work will significantly improve the robustness of segmentation models to domain shifts in real-world clinical situations where the absence of source and target data is the norm.

APPENDIX I

ESTIMATION OF THE SHAPE MOMENTS FROM ANATOMICAL KNOWLEDGE

We detail below the estimation of the relevant shape moments from anatomical knowledge for each application. For the 2D models $TTAS_2$, an estimation of the 2D class-ratio $\mathcal{R}$ of the anatomical structures is needed. For the 3D models $TTAS_3$, an estimation of the 3D class-ratio $\mathcal{R}$ and the 3D distance-to-centroid $\mathcal{D}$. We denote r_u, r_v, r_w the resolution values in the corresponding plane and Ω the cardinal size of the image or volume. Note that for a structure k, after obtaining the estimated size $sz^{(k)}$ in mm^2 (2D case) or the volume $V^{(k)}$ in mm^3 (3D case), the class-ratio (i.e. region proportion) $\mathcal{R}_k$ is calculated as:

$$\mathcal{R}_k = \begin{cases} \dfrac{sz^{(k)}}{r_u * r_v * \Omega} \text{ for } TTAS_2. \\ \\ \dfrac{V^{(k)}}{r_u * r_v * r_w * \Omega} \text{ for } TTAS_3. \end{cases} \tag{A I-1}$$

Similarly, the distance-to-centroid in pixel is derived from the estimated distance-to-centroid in mm $(d_u^{(k)}, d_v^{(k)}, d_w^{(k)})$:

$$\mathcal{D}_k = \begin{cases} (\dfrac{d_u^{(k)}}{r_u}, \dfrac{d_v^{(k)}}{r_v}) \text{ for } TTAS_2. \\ \\ (\dfrac{d_u^{(k)}}{r_u}, \dfrac{d_v^{(k)}}{r_v}, \dfrac{d_w^{(k)}}{r_w}) \text{ for } TTAS_3. \end{cases} \tag{A I-2}$$

For each coordinate (u, v, w), we derive an estimation of its corresponding distance-to-centroid $(\mathcal{D}_k(u), \mathcal{D}_k(v), \mathcal{D}_k(w))$, from the estimated radius $(R_k(u),...)$ in mm as follows:

$$\mathcal{D}_k(u) = \sqrt{(R_k(u) + 1) * (R_k(u) * 2 + 1)/6)}.$$

Table I-1 summarizes the estimations obtained for each structure.

NCI-ISBI13 Prostate volume and dimensions are widely monitored as they are biomarkers of prostate health. Reference volume $V_{prostate}$ and radius R_u, R_v, R_w were taken from Eri, Thomassen, Brennhovd & Håheim (2002), which measured them through planimetry. We then estimated the transverse surface dimension as: $sz_{prostate} = \frac{3V_{prostate}}{4R_w}$.

MMWHS [14]. Measuring and estimating 2D and 3D heart substructure dimensions is a well-studied problem in cardiology, as these values are the main biomarkers in the clinical assessment of heart diseases. Therefore, reference heart substructure dimensions are easily accessible in every plane. **LA** For the 2D case, we used an estimation of the Left Atrium size sz_{LA} from the measurements in Anderson, Horne & Pennell (2005) Table 1, taken at maximum volume (end-systole) in the 4-chamber view[15] For the 3D case, we used the estimation of the Left Atrium volume and radius from Pritchett, Jacobsen, Mahoney, Rodeheffer, Bailey & Redfield (2003); **LV and Myo** We use measurements from Støylen, Dalen & Molmen (2020), where left ventricular myocardial and cavity volumes are available at end-diastole (LVEDV and MVd respectively for the volumes) and end-systole (LVESV and MVs). We estimate the left ventricular volume V_{LV} as : $V_{LV} = \frac{LVEDV+LVESV}{2}$ and the myocardium volume as $V_{Myo} = \frac{MVd+MVs}{2}$. Similarly, the outer and inner LV diastolic length (LVLd and LVILd), the diastolic and systolic wall thicknesses (WTd and WTs), and the outer LV diastolic and systolic diameters (LVEDd and LVEDs) were used to compute the LV and Myo radii. For the 2D case, we obtained an estimation of sz_{LV} from O'Dell (2019); to derive an estimation of sz_{Myo}, we computed these two ratios: $r_{diastole} = \frac{LVEDV}{MVd}$; $r_{systole} = \frac{LVESV}{MVs}$ and estimated the average size in a coronal slice as : $sz_{Myo} = \frac{r_{diastole}+r_{systole}}{2} * sz_{LV}$. **AA** We used aortic diameters at proximal (p) and distal (d) levels as given in Aronberg, Glazer, Madsen & Sagel (1984), as well as the average AA length (l)

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

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

provided by the MMWHS organisers [16] to obtain estimation of the radii. Then, we derived an estimation of the average AA area in a coronal slice as: $sz_{AA} = \frac{p+d}{2} * l + \pi * (p/4)^2$, and of the 3D volume as: $V_{AA} = \frac{p+d}{2} * l^2 + \frac{4}{2}\pi * (p/4)^3$.

Table-A I-1 Estimated shape moments of structures in the target datasets

	Shape descriptors	**NCI-ISBI13** Prostate	**MMWHS** Myo	LA	LV	AA
2D	$\mathcal{R}_k$ (%)	4.68	6.76	7.62	6.85	5.65
3D	$\mathcal{R}_k$ (%)	3.10	2.18	1.52	1.43	1.21
	$\mathcal{D}_k(u)$ (vox)	22	33	22	25	16
	$\mathcal{D}_k(v)$ (vox)	25	30	25	23	32
	$\mathcal{D}_k(w)$ (vox)	3	34	15	23	16

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

APPENDIX II

LINK BETWEEN THE LOSS IN CHAPTER 3 AND MUTUAL INFORMATION MAXIMIZATION

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

$$\mathcal{I}(X;Y) = \mathbb{E}_Y\left[\log \mathbb{E}_X[p(Y \mid X)]\right] - \mathbb{E}_{X,Y}[\log p(Y \mid X)] \tag{A II-1}$$

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

$$\mathcal{I}(I_t;P_t) = \mathbb{E}_{P_t}\left[\log \mathbb{E}_{I_t}[p(P_t \mid I_t)]\right] - \mathbb{E}_{I_t,P_t}[\log p(P_t \mid I_t)] \tag{A II-2}$$

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

$$\begin{aligned}
\mathbb{E}_{P_t}\left[\log \mathbb{E}_{I_t}[p(P_t \mid I_t)]\right] &= \mathbb{E}_{P_t}\left[\log \hat{\tau}(t,.,\theta)\right] \\
&= \frac{1}{|\Omega_t|}\sum_{i\in\Omega_t}\sum_{k=1}^{K} p_t^k(i,\theta)\log\hat{\tau}(t,k,\theta) \\
&= \sum_{k=1}^{K}\hat{\tau}(t,k,\theta)\log\hat{\tau}(t,k,\theta) = -H\{\hat{\tau}(t,.,\theta)\}
\end{aligned} \tag{A II-3}$$

and :

$$
\begin{aligned}
-\mathbb{E}_{I_t,P_t}[\log p(P_t \mid I_t)] &= -\frac{1}{|\Omega_t|}\sum_{i\in\Omega_t}\sum_{k=1}^{K} p_t^k(i,\theta)\log p_t^k(i,\theta) \\
&= \frac{1}{|\Omega_t|}\sum_{i\in\Omega_t}\ell_{ent}(\mathbf{p}_t(i,\theta))\}
\end{aligned}
\tag{A II-4}
$$

The following identity follows:

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

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

$$
\mathcal{I}_\theta = \frac{1}{T}\sum_{t=1}^{T}\mathcal{I}(I_t;P_t) = \frac{1}{T}\sum_{t=1}^{T} -H\{\hat{\tau}(t,.,\theta)\} + \frac{1}{|\Omega_t|}\sum_{i\in\Omega_t}\ell_{ent}(\mathbf{p}_t(i,\theta)) \tag{A II-6}
$$

BIBLIOGRAPHY

Abbasi-Sureshjani, S., Raumanns, R., Michels, B., Schouten, G. & Cheplygina, V. (2020). Risk of Training Diagnostic Algorithms on Data with Demographic Bias. *MICCAI Workshop on Interpretable and Annotation-Efficient Learning*, pp. 183–192.

Acosta, J. N., Falcone, G. J., Rajpurkar, P. & Topol, E. J. (2022). Multimodal biomedical AI. *Nature Medicine*, 28(9), 1773–1784.

Anderson, J., Horne, B. & Pennell, D. (2005). Atrial Dimensions in Health and Left Ventricular Disease Using Cardiovascular Magnetic Resonance. *Journal of the Society for Cardiovascular Magnetic Resonance*, 7, 671–5.

Aronberg, D., Glazer, H., Madsen, K. & Sagel, S. (1984). Normal thoracic aortic diameters by computed tomography. *Computer Assisted Tomography*, 8(2), 247–250.

Ashukha, A., Lyzhov, A., Molchanov, D. & Vetrov, D. (2020). Pitfalls of In-Domain Uncertainty Estimation and Ensembling in Deep Learning. *International Conference on Learning Representations (ICLR)*.

Badgeley, M. A., Zech, J. R., Oakden-Rayner, L., Glicksberg, B. S., Liu, M., Gale, W. et al. (2019). Deep learning predicts hip fracture using confounding patient and healthcare variables. *NPJ Digital Medicine*, 2(1), 1–10.

Bai, J., Miri, M. S., Liu, Y., Saha, P. K., Garvin, M. K. & Wu, X. (2014). Graph-based optimal multi-surface segmentation with a star-shaped prior: Application to the segmentation of the optic disc and cup. *International Symposium on Biomedical Imaging (ISBI)*, pp. 525–528.

Bai, W., Oktay, O., Sinclair, M., Suzuki, H., Rajchl, M., Tarroni, G. et al. (2017). Semi-supervised learning for network-based cardiac MR image segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 253–260.

Bateson, M., Dolz, J., Kervadec, H., Lombaert, H. & Ben Ayed, I. (2021). Constrained Domain Adaptation for Image Segmentation. *IEEE Transactions on Medical Imaging (TMI)*, 40(7), 326–334.

Bateson, M., Dolz, J., Kervadec, H., Lombaert, H. & Ayed, I. B. (2019). Constrained domain adaptation for segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 326–334.

Bateson, M., Kervadec, H., Dolz, J., Lombaert, H. & Ben Ayed, I. (2020). Source-Relaxed Domain Adaptation for Image Segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 490–499.

Bateson, M., Kervadec, H., Dolz, J., Lombaert, H. & Ben Ayed, I. (2022a). Source-free domain adaptation for image segmentation. *Medical Image Analysis*, 82, 102617.

Bateson, M., Lombaert, H. & Ayed, I. B. (2022b). Test-Time Adaptation with Shape Moments for Image Segmentation. arXiv preprint 2205.07983.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A. et al. (2010). A theory of learning from different domains. *Machine learning*, 79(1-2), 151–175.

Benaim, S. & Wolf, L. (2018). One-Shot Unsupervised Cross Domain Translation. *Advances in Neural Information Processing Systems (NIPS)*, pp. 2108–2118.

Berry, J. L., Moran, J. M., Berg, W. S. & Steffee, A. D. (1987). A morphometric study of human lumbar and selected thoracic vertebrae. *Spine*, 12(4), 362–367.

Bertsekas, D. P. (1995). *Nonlinear Programming*. Athena Scientific, Belmont, MA.

Bertsekas, D. P. (1976). Multiplier methods: A survey. *Automatica*, 12(2), 133–145.

Bian, C., Yuan, C., Wang, J., Li, M., Yang, X., Yu, S. et al. (2020). Uncertainty-aware domain alignment for anatomical structure segmentation. *Medical Image Analysis*, 64, 101732.

Billot, B., Greve, D. N., Van Leemput, K., Fischl, B., Iglesias, J. E. & Dalca, A. (2020). A Learning Strategy for Contrast-agnostic MRI Segmentation. *Medical Imaging with Deep Learning*, pp. 75–93.

Billot, B., Greve, D. N., Puonti, O., Thielscher, A., Van Leemput, K., Fischl, B. et al. (2021). SynthSeg: Domain Randomisation for Segmentation of Brain MRI Scans of any Contrast and Resolution. arXiv preprint 2107.09559.

Blake, A., Kohli, P. & Rother, C. (2011). *Markov random fields for vision and image processing*. Mit Press.

Bohlender, S., Oksuz, I. & Mukhopadhyay, A. (2021). A survey on shape-constraint deep learning for medical image segmentation. arXiv preprint 2101.07721.

Bottou, L. & Bousquet, O. (2007). The tradeoffs of large scale learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 20, 161–168.

Boudiaf, M., Kervadec, H., Masud, Z. I., Piantanida, P., Ben Ayed, I. & Dolz, J. (2021). Few-Shot Segmentation Without Meta-Learning: A Good Transductive Inference Is All You Need? *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13979–13988.

Boudiaf, M., Mueller, R., Ben Ayed, I. & Bertinetto, L. (2022). Parameter-free Online Test-time Adaptation. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8344–8353.

Bousmalis, K., Silberman, N., Dohan, D., Erhan, D. & Krishnan, D. (2017). Unsupervised Pixel-Level Domain Adaptation with Generative Adversarial Networks. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 95–104.

Bousmalis, K., Trigeorgis, G., Silberman, N., Krishnan, D. & Erhan, D. (2016). Domain separation networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 29, 343–351.

Boyd, S. & Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

Boykov, Y., Isack, H., Olsson, C. & Ben Ayed, I. (2015). Volumetric Bias in Segmentation and Reconstruction: Secrets and Solutions. *International Conference on Computer Vision (ICCV)*, pp. 1769–1777.

Bucci, S., Loghmani, M. R. & Tommasi, T. (2020). On the effectiveness of image rotation for open set domain adaptation. *European Conference on Computer Vision (ECCV)*, pp. 422–438.

Carlucci, F. M., D'Innocente, A., Bucci, S., Caputo, B. & Tommasi, T. (2019). Domain generalization by solving jigsaw puzzles. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2229–2238.

Chang, W.-L., Wang, H.-P., Peng, W.-H. & Chiu, W.-C. (2019a). All About Structure: Adapting Structural Information Across Domains for Boosting Semantic Segmentation. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1900–1909.

Chang, W.-G., You, T., Seo, S., Kwak, S. & Han, B. (2019b). Domain-specific batch normalization for unsupervised domain adaptation. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7354–7362.

Chen, C., Qin, C., Qiu, H., Tarroni, G., Duan, J., Bai, W. et al. (2020a). Deep Learning for Cardiac Image Segmentation: A Review. *Frontiers in Cardiovascular Medicine*, 7, 1–1.

Chen, C., Dou, Q., Chen, H. & Heng, P.-A. (2018a). Semantic-Aware Generative Adversarial Nets for Unsupervised Domain Adaptation in Chest X-Ray Segmentation. *International Conference on Machine Learning in Medical Imaging (MIDL)*, pp. 143–151.

Chen, C., Dou, Q., Chen, H., Qin, J. & Heng, P. A. (2020b). Unsupervised Bidirectional Cross-Modality Adaptation via Deeply Synergistic Image and Feature Alignment for Medical Image Segmentation. *IEEE Transactions on Medical Imaging (TMI)*, 39(7), 2494–2505.

Chen, C., Liu, Q., Jin, Y., Dou, Q. & Heng, P.-A. (2021a). Source-Free Domain Adaptive Fundus Image Segmentation with Denoised Pseudo-Labeling. *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 225–235.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K. & Yuille, A. L. (2015). Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. *International Conference on Learning Representations (ICLR)*.

Chen, L.-C., Papandreou, G., Schroff, F. & Adam, H. (2017a). Rethinking atrous convolution for semantic image segmentation. arXiv preprint 1706.05587.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K. & Yuille, A. L. (2018b). DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 40(4), 834–848.

Chen, R. J., Lu, M. Y., Weng, W.-H., Chen, T. Y., Williamson, D. F., Manz, T. et al. (2021b). Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. *International Conference on Computer Vision (ICCV)*, pp. 4015–4025.

Chen, X. & Konukoglu, E. (2018). Unsupervised Detection of Lesions in Brain MRI using constrained adversarial auto-encoders. *International Conference on Medical Imaging with Deep Learning (MIDL)*.

Chen, Y.-H., Chen, W.-Y., Chen, Y.-T., Tsai, B.-C., Wang, Y.-C. F. & Sun, M. (2017b). No More Discrimination: Cross City Adaptation of Road Scene Segmenters. *International Conference on Computer Vision (ICCV)*, pp. 1992–2001.

Chen, Y., Wei, C., Kumar, A. & Ma, T. (2020c). Self-training avoids using spurious features under domain shift. *Advances in Neural Information Processing Systems (NeurIPS)*, 33, 21061–21071.

Chen, Y., Li, W. & Van Gool, L. (2018c). Road: Reality oriented adaptation for semantic segmentation of urban scenes. *Computer Vision and Pattern Recognition*, pp. 7892–7901.

Cheplygina, V., de Bruijne, M. & Pluim, J. P. W. (2019). Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Medical Image Analysis*, 54, 280–296.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T. & Ronneberger, O. (2016). 3D U-Net: learning dense volumetric segmentation from sparse annotation. *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 424–432.

Cohen, J. P., Luck, M. & Honari, S. (2018). Distribution matching losses can hallucinate features in medical image translation. *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 529–536.

Crammer, K., Kearns, M. & Wortman, J. (2008). Learning from Multiple Sources. *Journal of Machine Learning Research*, 9(57), 1757–1774.

Cremers, D., Osher, S. J. & Soatto, S. (2006). Kernel Density Estimation and Intrinsic Alignment for Shape Priors in Level Set Segmentation. *International Journal of Computer Vision (IJCV)*, 69(3), 335–351.

Cui, H., Wang, X., Lin, W., Zhou, J., Eberl, S., Feng, D. et al. (2016). Primary lung tumor segmentation from PET–CT volumes with spatial–topological constraint. *International Journal of Computer-Assisted Radiology and Surgery*, 11(1), 19–29.

Czipczer, V. & Manno-Kovacs, A. (2022). Adaptable volumetric liver segmentation model for CT images using region-based features and convolutional neural network. *Neurocomputing*, 505, 388–401.

Dai, J., He, K. & Sun, J. (2015). BoxSup: Exploiting Bounding Boxes to Supervise Convolutional Networks for Semantic Segmentation. *International Conference on Computer Vision (ICCV)*, pp. 1635–1643.

Davnall, F., Yip, C. S. P., Ljungqvist, G., Selmi, M., Ng, F., Sanghera, B. et al. (2012). Assessment of tumor heterogeneity: an emerging imaging tool for clinical practice? *Insights into Imaging*, 3(6), 573–589.

DeGrave, A. J., Janizek, J. D. & Lee, S.-I. (2021). AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3(7), 610–619.

Diaz-Pinto, A., Ravikumar, N., Attar, R., Suinesiaputra, A., Zhao, Y., Levelt, E. et al. (2022). Predicting myocardial infarction through retinal scans and minimal personal information. *Nature Machine Intelligence*, 4(1), 55–61.

Ding, Z., Han, X. & Niethammer, M. (2020). Votenet+: An improved deep learning label fusion method for multi-atlas segmentation. *International Symposium on Biomedical Imaging (ISBI)*, pp. 363–367.

Doersch, C., Gupta, A. & Efros, A. A. (2015). Unsupervised visual representation learning by context prediction. *International Conference on Computer Vision (ICCV)*, pp. 1422–1430.

Dolz, J., Desrosiers, C. & Ayed, I. B. (2017). 3D fully convolutional networks for subcortical segmentation in MRI: A large-scale study. *NeuroImage*, 170, -.

Dorent, R., Joutard, S., Shapey, J., Bisdas, S., Kitchen, N., Bradford, R. et al. (2020). Scribble-based Domain Adaptation via Co-segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 479–489.

Dorent, R., Joutard, S., Shapey, J., Kujawa, A., Modat, M., Ourselin, S. et al. (2021). Inter extreme points geodesics for end-to-end weakly supervised image segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 615–624.

Dou, Q., Ouyang, C., Chen, C., Chen, H., Glocker, B., Zhuang, X. et al. (2019). PnP-AdaNet: Plug-and-Play Adversarial Domain Adaptation Network at Unpaired Cross-Modality Cardiac Segmentation. *IEEE Access*, 7, 99065–99076.

Dusenberry, M. W., Tran, D., Choi, E., Kemp, J., Nixon, J., Jerfel, G. et al. (2020). Analyzing the role of model uncertainty for electronic health records. *ACM Conference on Health, Inference, and Learning*, pp. 204–213.

El Jurdi, R., Petitjean, C., Honeine, P., Cheplygina, V. & Abdallah, F. (2021). High-level prior-based loss functions for medical image segmentation: A survey. *Computer Vision and Image Understanding*, 1–1.

Elyan, E., Vuttipittayamongkol, P., Johnston, P., Martin, K., McPherson, K., Jayne, C. et al. (2022). Computer vision and machine learning for medical image analysis: recent advances, challenges, and way forward. *Artificial Intelligence Surgery*, 2, 1–1.

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

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M. et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639), 115–118.

Esteva, A., Chou, K., Yeung, S., Naik, N., Madani, A., Mottaghi, A. et al. (2021). Deep learning-enabled medical computer vision. *NPJ Digital Medicine*, 4(1), 5.

Ganin, Y. & Lempitsky, V. (2015). Unsupervised Domain Adaptation by Backpropagation. *International Conference on Machine Learning (ICML)*, pp. 1180–1189.

Ganin, Y. et al. (2016). Domain-Adversarial Training of Neural Networks. *Journal of Machine Learning Research (JMLR)*, 17(1), 2096–2030.

Ghafoorian, M., Karssemeijer, N., Heskes, T., van Uder, I. W. M., de Leeuw, F. E., Marchiori, E. et al. (2016). Non-uniform patch sampling with deep convolutional neural networks for white matter hyperintensity segmentation. *International Symposium on Biomedical Imaging (ISBI)*, pp. 1414–1417.

Ghifary, M., Kleijn, W. B., Zhang, M. & Balduzzi, D. (2015). Domain generalization for object recognition with multi-task autoencoders. *International Conference on Computer Vision (ICCV)*, pp. 2551–2559.

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

Gholami, A. et al. (2018). A Novel Domain Adaptation Framework for Medical Image Segmentation. *MICCAI Brainlesion Workshop*, pp. 289–298.

Goodfellow, I. J. et al. (2014). Generative Adversarial Nets. *Advances in Neural Information Processing Systems (NIPS)*, pp. 2672–2680.

Gordillo, N., Montseny, E. & Sobrevilla, P. (2013). State of the art survey on MRI brain tumor segmentation. *Magnetic Resonance Imaging*, 31(8), 1426–1438.

Grandvalet, Y. & Bengio, Y. (2004). Semi-Supervised Learning by Entropy Minimization. *Advances in Neural Information Processing Systems (NIPS)*.

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

Gupta, S., Hu, X., Kaan, J., Jin, M., Mpoy, M., Chung, K. et al. (2022). Learning Topological Interactions for Multi-Class Medical Image Segmentation. *European Conference on Computer Vision (ECCV)*, pp. 701–718.

Hardt, M., Recht, B. & Singer, Y. (2016). Train faster, generalize better: Stability of stochastic gradient descent. *International Conference on Machine Learning (ICML)*, pp. 1225–1234.

He, F. S., Liu, Y., Schwing, A. G. & Peng, J. (2017). Learning to Play in a Day: Faster Deep Reinforcement Learning by Optimality Tightening. *International Conference on Learning Representations (ICLR)*, pp. 1–13.

He, K., Zhang, X., Ren, S. & Sun, J. (2016). Deep residual learning for image recognition. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778.

He, Y., Carass, A., Zuo, L., Dewey, B. E. & Prince, J. L. (2020). Self Domain Adapted Network. *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 437–446.

He, Y., Carass, A., Zuo, L., Dewey, B. E. & Prince, J. L. (2021). Autoencoder based self-supervised test-time adaptation for medical image analysis. *Medical Image Analysis*, 72, 102136.

HeartFlowNXT. (2017). HeartFlow Analysis of Coronary Blood Flow Using Coronary CT Angiography–Study Results–ClinicalTrials.gov. https://clinicaltrials.gov/ct2/show/results/NCT01757678. https://clinicaltrials.gov/ct2/show/results/NCT01757678.

Heiliger, L., Sekuboyina, A., Menze, B., Egger, J. & Kleesiek, J. (2022). Beyond Medical Imaging - A Review of Multimodal Deep Learning in Radiology. techrXiv 10.36227/techrxiv.19103432.v1.

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E. et al. (2021). The many faces of robustness: A critical analysis of out-of-distribution generalization. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8340–8349.

Hoffman, J., Tzeng, E., Park, T., Zhu, J.-Y., Isola, P., Saenko, K. et al. (2018). Cycada: Cycle-consistent adversarial domain adaptation. *International Conference on Machine Learning (ICML)*, 80, 1989–1998.

Hojjatoleslami, A. & Avanaki, M. R. N. (2012). OCT skin image enhancement through attenuation compensation. *Applied Optics*, 51(21), 4927–4935.

Hong, W., Wang, Z., Yang, M. & Yuan, J. (2018). Conditional Generative Adversarial Network for Structured Domain Adaptation. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1335–1344.

Hu, J., Zhong, H., Yang, F., Gong, S., Wu, G. & Yan, J. (2022). Learning Unbiased Transferability for Domain Adaptation by Uncertainty Modeling. *European Conference on Computer Vision (ECCV)*, pp. 223–241.

Hu, M., Song, T., Gu, Y., Luo, X., Chen, J., Chen, Y. et al. (2021a). Fully Test-Time Adaptation for Image Segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 251–260.

Hu, X., Uzunbas, G., Chen, S., Wang, R., Shah, A., Nevatia, R. et al. (2021b). Mixnorm: Test-time adaptation through online normalization estimation. arXiv preprint 2110.11478.

Huang, H., Zheng, H., Lin, L., Cai, M., Hu, H., Zhang, Q. et al. (2021). Medical Image Segmentation With Deep Atlas Prior. *IEEE Transactions on Medical Imaging (TMI)*, 40(12), 3519–3530.

Huo, Y., Xu, Z., Bao, S., Assad, A., Abramson, R. G. & Landman, B. A. (2018). Adversarial synthesis learning enables segmentation without target modality ground truth. *International Symposium on Biomedical Imaging (ISBI)*, pp. 1217–1220.

Huo, Y. et al. (2019). SynSeg-Net: Synthetic Segmentation Without Target Modality Ground Truth. *IEEE Transactions on Medical Imaging (TMI)*, 38(4), 1016–1025.

Ioffe, S. & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *International Conference on Machine Learning (ICML)*, pp. 448–456.

Isack, H., Gorelick, L., Ng, K., Veksler, O. & Boykov, Y. (2018). K-convexity shape priors for segmentation. *European Conference on Computer Vision (ECCV)*, pp. 36–51.

Jabi, M., Pedersoli, M., Mitiche, A. & Ayed, I. B. (2021). Deep Clustering: On the Link Between Discriminative Models and K-Means. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(6), 1887–1896.

Javanmardi, M. & Tasdizen, T. (2018). Domain adaptation for biomedical image segmentation using adversarial training. *International Symposium on Biomedical Imaging (ISBI)*, pp. 554–558.

Jia, H., Cai, W., Huang, H. & Xia, Y. (2022). Learning multi-scale synergic discriminative features for prostate image segmentation. *Pattern Recognition*, 126, 108556.

Jia, Z., Huang, X., Eric, I., Chang, C. & Xu, Y. (2017). Constrained deep weak supervision for histopathology image segmentation. *IEEE Transactions on Medical Imaging (TMI)*, 36(11), 2376–2388.

Jog, A., Hoopes, A., Greve, D. N., Van Leemput, K. & Fischl, B. (2019). PSACNN: Pulse sequence adaptive fast whole brain segmentation. *NeuroImage*, 199, 553–569.

Kaissis, G. A., Makowski, M. R., Rückert, D. & Braren, R. F. (2020). Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, 2(6), 305–311.

Kamnitsas, K. et al. (2017). Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. *International Conference on Information Processing in Medical Imaging (IPMI)*, 10265, 597–609.

Karani, N., Erdil, E., Chaitanya, K. & Konukoglu, E. (2021). Test-time adaptable neural networks for robust medical image segmentation. *Medical Image Analysis*, 68, 101907.

Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G. & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine*, 17(1), 195.

Kervadec, H., Dolz, J., Granger, É. & Ben Ayed, I. (2019a). Curriculum Semi-supervised Segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 568–576.

Kervadec, H., Dolz, J., Tang, M., Granger, E., Boykov, Y. & Ben Ayed, I. (2019b). Constrained-CNN losses for weakly supervised segmentation. *Medical Image Analysis*, 54, 88–99.

Kervadec, H., Dolz, J., Wang, S., Granger, E. & Ben Ayed, I. (2020). Bounding boxes for weakly supervised segmentation: Global constraints get close to full supervision. *International Conference on Medical Imaging with Deep Learning (MIDL)*, pp. 365–381.

Kervadec, H., Bahig, H., Létourneau-Guillon, L., Dolz, J. & Ben Ayed, I. (2021). Beyond pixel-wise supervision for segmentation: A few global shape descriptors might be surprisingly good! *International Conference on Medical Imaging with Deep Learning (MIDL)*, 143, 354–368.

Kervadec, H., Dolz, J., Yuan, J., Desrosiers, C., Granger, E. & Ben Ayed, I. (2022). Constrained deep networks: Lagrangian optimization via Log-barrier extensions. *European Signal Processing Conference*, pp. 962–966.

Khalili, N., Turk, E., Zreik, M., Viergever, M. A., Benders, M. J. N. L. & Išgum, I. (2019). Generative Adversarial Network for Segmentation of Motion Affected Neonatal Brain MRI. *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 320–328.

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

Kingma, D. & Ba, J. (2014). Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations (ICLR)*.

Klodt, M. & Cremers, D. (2011a). segmentation with moment constraints. *International Conference on Computer Vision (ICCV)*, pp. 2236–2243.

Klodt, M. & Cremers, D. (2011b). A Convex Framework for Image Segmentation with Moment Constraints. *International Conference on Computer Vision (ICCV)*, pp. 2236–2243.

Kolesnikov, A., Zhai, X. & Beyer, L. (2019). Revisiting self-supervised visual representation learning. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1920–1929.

Kowarsch, F., Weijler, L., Wödlinger, M., Reiter, M., Maurer-Granofszky, M., Schumich, A. et al. (2022). Towards Self-explainable Transformers for Cell Classification in Flow Cytometry Data. *MICCAI Workshop on Interpretability of Machine Intelligence in Medical Image Computing*, pp. 22–32.

Krähenbühl, P. & Koltun, V. (2011). Efficient inference in fully connected crfs with gaussian edge potentials. *Advances in Neural Information Processing Systems (NIPS)*, 24, 109–117.

Krause, A., Perona, P. & Gomes, R. (2010). Discriminative Clustering by Regularized Information Maximization. *Advances in Neural Information Processing Systems (NIPS)*, 23, 775–783.

Krizhevsky, A., Sutskever, I. & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 25, 1097–1105.

Kundu, J. N., Kulkarni, A., Singh, A., Jampani, V. & Babu, R. V. (2021). Generalize then adapt: Source-free domain adaptive semantic segmentation. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7046–7056.

Lara, M. A. R., Echeveste, R. & Ferrante, E. (2022). Addressing fairness in artificial intelligence for medical imaging. *Nature Communications*, 13(1), 4581.

Lavin, A., Gilligan-Lee, C. M., Visnjic, A., Ganju, S., Newman, D., Ganguly, S. et al. (2022). Technology readiness levels for machine learning systems. *Nature Communications*, 13(1), 6039.

Le, Q. V., Monga, R., Devin, M., Corrado, G., Chen, K., Ranzato, M. et al. (2013). Building high-level features using large scale unsupervised learning. *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8595–8598.

LeCun, Y., Bengio, Y. & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444.

Li, Y., Wang, N., Shi, J., Liu, J. & Hou, X. (2016). Revisiting batch normalization for practical domain adaptation. arXiv preprint 1603.04779.

Li, Y., Wang, N., Shi, J., Hou, X. & Liu, J. (2018). Adaptive batch normalization for practical domain adaptation. *Pattern Recognition*, 80, 109–117.

Li, Y., Yuan, L. & Vasconcelos, N. (2019). Bidirectional Learning for Domain Adaptation of Semantic Segmentation. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6929–6938.

Liang, J., He, R., Sun, Z. & Tan, T. (2019). Exploring uncertainty in pseudo-label guided unsupervised domain adaptation. *Pattern Recognition*, 96, 106996.

Liang, J., Hu, D. & Feng, J. (2020). Do We Really Need to Access the Source Data? Source Hypothesis Transfer for Unsupervised Domain Adaptation. *International Conference on Machine Learning (ICML)*, 119, 6028–6039.

Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollár, P. (2018). Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2999-3007.

Litjens, G. et al. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42, 60–88.

Liu, B., Dolz, J., Galdran, A., Kobbi, R. & Ayed, I. B. (2021a). The hidden label-marginal biases of segmentation losses. arXiv preprint 2104.08717.

Liu, B., Ben Ayed, I., Galdran, A. & Dolz, J. (2022). The Devil is in the Margin: Margin-based Label Smoothing for Network Calibration. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 80–88.

Liu, M.-Y. & Tuzel, O. (2016). Coupled generative adversarial networks. *Advances in Neural Information Processing Systems (NIPS)*, 29, 1–1.

Liu, Q., Dou, Q. & Heng, P.-A. (2020). Shape-Aware Meta-learning for Generalizing Prostate MRI Segmentation to Unseen Domains. *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 475–485.

Liu, Q., Chen, C., Qin, J., Dou, Q. & Heng, P. (2021b). FedDG: Federated Domain Generalization on Medical Image Segmentation via Episodic Learning in Continuous Frequency Space. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1013–1023.

Liu, X., Xing, F., Yang, C., El Fakhri, G. & Woo, J. (2021c). Adapting Off-the-Shelf Source Segmenter for Target Medical Image Segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 549–559.

Liu, Y., Zhang, W. & Wang, J. (2021d). Source-Free Domain Adaptation for Semantic Segmentation. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1215–1224.

Long, J., Shelhamer, E. & Darrell, T. (2015a). Fully convolutional networks for semantic segmentation. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3431–3440.

Long, M., Cao, Y., Wang, J. & Jordan, M. I. (2015b). Learning transferable features with deep adaptation networks. *International Conference on Machine Learning (ICML)*, pp. 97–105.

Long, M., Zhu, H., Wang, J. & Jordan, M. I. (2016). Unsupervised Domain Adaptation with Residual Transfer Networks. *Advances in Neural Information Processing Systems (NIPS)*, pp. 136–144.

Lorenzo-Valdés, M., Sanchez-Ortiz, G. I., Mohiaddin, R. & Rueckert, D. (2002). Atlas-based segmentation and tracking of 3D cardiac MR images using non-rigid registration. *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 642–650.

Luo, Y., Zheng, L., Guan, T., Yu, J. & Yang, Y. (2019). Taking a Closer Look at Domain Shift: Category-Level Adversaries for Semantics Consistent Domain Adaptation. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2507–2516.

Márquez-Neila, P. et al. (2017). Imposing Hard Constraints on Deep Networks: Promises and Limitations. *Computer Vision and Pattern Recognition Workshop on Negative Results*, pp. 1–9.

Milletari, F., Navab, N. & Ahmadi, S.-A. (2016). V-Net: Fully convolutional neural networks for volumetric medical image segmentation. *3D Vision (3DV), Fourth International Conference on*, pp. 565–571.

Mirikharaji, Z. & Hamarneh, G. (2018). Star shape prior in fully convolutional networks for skin lesion segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 737–745.

Moon, J. H., Lee, H., Shin, W., Kim, Y.-H. & Choi, E. (2022). Multi-Modal Understanding and Generation for Medical Images and Text via Vision-Language Pre-Training. *IEEE Journal of Biomedical and Health Informatics*, 26(12), 6070–6080.

Morerio, P., Cavazza, J. & Murino, V. (2018). Minimal-Entropy Correlation Alignment for Unsupervised Deep Domain Adaptation. *International Conference on Learning Representations (ICLR)*.

Motiian, S., Piccirilli, M., Adjeroh, D. A. & Doretto, G. (2017). Unified Deep Supervised Domain Adaptation and Generalization. *International Conference on Computer Vision (ICCV)*, pp. 5715–5725.

Mukhoti, J., Kulharia, V., Sanyal, A., Golodetz, S., Torr, P. & Dokania, P. (2020). Calibrating deep neural networks using focal loss. *Advances in Neural Information Processing Systems (NeurIPS)*, 33, 15288–15299.

Mummadi, C. K., Hutmacher, R., Rambach, K., Levinkov, E., Brox, T. & Metzen, J. H. (2021). Test-time adaptation to distribution shift by confidence maximization and input transformation. arXiv preprint 2106.14999.

Nado, Z., Padhy, S., Sculley, D., D'Amour, A., Lakshminarayanan, B. & Snoek, J. (2020). Evaluating prediction-time batch normalization for robustness under covariate shift. arXiv preprint 2006.10963.

Nath Kundu, J., Venkat, N., Rahul, M. V. & Venkatesh Babu, R. (2020). Universal Source-Free Domain Adaptation. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4543–4552.

Nosrati, M. S. & Hamarneh, G. (2016). Incorporating prior knowledge in medical image segmentation: a survey. arXiv preprint 1607.01092.

Obermeyer, Z. & Topol, E. J. (2021). Artificial intelligence, bias, and patients' perspectives. *The Lancet*, 397(10289), 2038.

O'Dell, W. G. (2019). Accuracy of Left Ventricular Cavity Volume and Ejection Fraction for Conventional Estimation Methods and 3D Surface Fitting. *Journal of the American Heart Association*, 8(6), e009124.

Oksuz, I., Mukhopadhyay, A., Dharmakumar, R. & Tsaftaris, S. A. (2017). Unsupervised Myocardial Segmentation for Cardiac BOLD. *IEEE Transactions on Medical Imaging (TMI)*, 36(11), 2228–2238.

Opbroek, A., Ikram, M., Vernooij, M. & de Bruijne, M. (2014). Transfer Learning Improves Supervised Image Segmentation Across Imaging Protocols. *IEEE Transactions on Medical Imaging (TMI)*, 34, 1018–1030.

Orbes-Arteaga, M., Varsavsky, T., Sudre, C. H., Eaton-Rosen, Z., Haddow, L. J., Sørensen, L. et al. (2019). Multi-domain adaptation in brain MRI through paired consistency and adversarial learning. In *MICCAI Workshop on Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data* (pp. 54–62). Springer.

Ouyang, C., Chen, C., Li, S., Li, Z., Qin, C., Bai, W. et al. (2022). Causality-inspired single-source domain generalization for medical image segmentation. *IEEE Transactions on Medical Imaging (TMI)*, 1–1.

Ouyang, X., Xue, Z., Zhan, Y., Zhou, X. S., Wang, Q., Zhou, Y. et al. (2019). Weakly supervised segmentation framework with uncertainty: A study on pneumothorax segmentation in chest x-ray. *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 613–621.

Pan, S. J. & Yang, Q. (2010). A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 22(10), 1345–1359.

Paszke, A., Chaurasia, A., Kim, S. & Culurciello, E. (2016). Enet: A deep neural network architecture for real-time semantic segmentation. arXiv preprint 1606.02147.

Patel, G. & Dolz, J. (2022). Weakly supervised segmentation with cross-modality equivariant constraints. *Medical Image Analysis*, 77, 102374.

Pathak, D., Krähenbühl, P. & Darrell, T. (2015). Constrained Convolutional Neural Networks for Weakly Supervised Segmentation. *International Conference on Computer Vision (ICCV)*, pp. 1796–1804.

Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T. & Efros, A. A. (2016). Context encoders: Feature learning by inpainting. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2536–2544.

Paul, S., Tsai, Y.-H., Schulter, S., Roy-Chowdhury, A. K. & Chandraker, M. (2020). Domain Adaptive Semantic Segmentation Using Weak Labels. *European Conference on Computer Vision (ECCV)*, pp. 571–587.

Pei, Z., Cao, Z., Long, M. & Wang, J. (2018). Multi-Adversarial Domain Adaptation. *AAAI Conference on Artificial Intelligence*.

Pinheiro, P. H. O. & Collobert, R. (2014). From image-level to pixel-level labeling with Convolutional Networks. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1713–1721.

Poplin, R., Varadarajan, A. V., Blumer, K., Liu, Y., McConnell, M. V., Corrado, G. S. et al. (2018). Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nature Biomedical Engineering*, 2(3), 158–164.

Pritchett, A. M., Jacobsen, S. J., Mahoney, D. W., Rodeheffer, R. J., Bailey, K. R. & Redfield, M. M. (2003). Left atrial volume as an index ofleft atrial size: a population-based study. *Journal of the American College of Cardiology*, 41(6), 1036–1043.

Puyol-Antón, E., Ruijsink, B., Piechnik, S. K., Neubauer, S., Petersen, S. E., Razavi, R. et al. (2021). Fairness in Cardiac MR Image Analysis: An Investigation of Bias Due to Data Imbalance in Deep Learning Based Segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)* (pp. 413–423). Springer.

Rajchl, M., Lee, M. C., Oktay, O., Kamnitsas, K., Passerat-Palmbach, J., Bai, W. et al. (2016). Deepcut: Object segmentation from bounding box annotations using convolutional neural networks. *IEEE Transactions on Medical Imaging (TMI)*, 36(2), 674–683.

Reddy, C., Gopinath, K. & Lombaert, H. (2019). Brain tumor segmentation using topological loss in convolutional networks. *International Conference on Medical Imaging with Deep Learning (MIDL)*.

Reinke, A., Eisenmann, M., Tizabi, M. D., Sudre, C. H., Rädsch, T., Antonelli, M. et al. (2021). Common limitations of image processing metrics: A picture story.

Ren, J. et al. (2018). Adversarial domain adaptation for classification of prostate histopathology whole-slide images. *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 201–209.

Rim, T. H., Lee, C. J., Tham, Y.-C., Cheung, N., Yu, M., Lee, G. et al. (2021). Deep-learning-based cardiovascular risk stratification using coronary artery calcium scores predicted from retinal photographs. *The Lancet Digital Health*, 3(5), e306–e316.

Ronneberger, O., Fischer, P. & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 234–241.

Rozantsev, A., Salzmann, M. & Fua, P. (2018). Beyond sharing weights for deep domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 41(4), 801–814.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S. et al. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3), 211–252.

Russo, P., Carlucci, F. M., Tommasi, T. & Caputo, B. (2018). From Source to Target and Back: Symmetric Bi-Directional Adaptive GAN. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8099–8108.

Sabanayagam, C., Xu, D., Ting, D., Nusinovici, S., Banu, R., Hamzah, H. et al. (2020). A deep learning algorithm to detect chronic kidney disease from retinal photographs in community-based populations. *The Lancet Digital Health*, 2, e295–e302.

Sangalli, S., Erdil, E., Hötker, A., Donati, O. & Konukoglu, E. (2021). Constrained optimization to train neural networks on critical and under-represented classes. *Advances in Neural Information Processing Systems (NeurIPS)*, 34, 25400–25411.

Sankaranarayanan, S., Balaji, Y., Castillo, C. & Chellappa, R. (2018). Generate to Adapt: Aligning Domains Using Generative Adversarial Networks. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8503–8512.

Saporta, A., Gui, X., Agrawal, A., Pareek, A., Truong, S. Q. H., Nguyen, C. D. T. et al. (2022). Benchmarking saliency methods for chest X-ray interpretation. *Nature Machine Intelligence*, 4(10), 867–878.

Schmidt, F. R. & Boykov, Y. (2012). Hausdorff Distance Constraint for Multi-surface Segmentation. *European Conference on Computer Vision (ECCV)*, pp. 598–611.

Schrouff, J., Harris, N., Koyejo, O., Alabdulmohsin, I., Schnider, E., Opsahl-Ong, K. et al. (2022). Maintaining fairness across distribution shift: do we have viable solutions for real-world applications? arXiv preprint 2202.01034.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *International Conference on Computer Vision (ICCV)*, pp. 618–626.

Shu, R., Bui, H. H., Narui, H. & Ermon, S. (2018). A DirtT-T approach to unsupervised domain adaptation. *International Conference on Learning Representations (ICLR)*.

Shuvaev, S., Lazutkin, A., Kiryanov, R., Anokhin, K., Enikolopov, G. & Koulakov, A. A. (2022). Spatiotemporal 3D image registration for mesoscale studies of brain development. *Scientific Reports*, 12(1), 3648.

Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C. A. et al. (2020). Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in Neural Information Processing Systems (NeurIPS)*, 33, 596–608.

Soni, N., Priya, S. & Bathla, G. (2019). Texture Analysis in Cerebral Gliomas: A Review of the Literature. *American Journal of Neuroradiology*, 40(6), 928–934.

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

Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S. & Cardoso, M. J. (2017). Generalised Dice overlap as a deep learning loss function for highly unbalanced segmentations. In *MICCAI Workshop on Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support* (pp. 240–248). Springer.

Sun, J., Darbehani, F., Zaidi, M. & Wang, B. (2020a). Saunet: Shape attentive u-net for interpretable medical image segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 797–806.

Sun, Y., Wang, X., Liu, Z., Miller, J., Efros, A. & Hardt, M. (2020b). Test-Time Training with Self-Supervision for Generalization under Distribution Shifts. *International Conference on Machine Learning (ICML)*, 119, 9229–9248.

Tajbakhsh, N., Jeyaseelan, L., Li, Q., Chiang, J. N., Wu, Z. & Ding, X. (2020). Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. *Medical Image Analysis*, 63, 101693.

Tang, M., Djelouah, A., Perazzi, F., Boykov, Y. & Schroers, C. (2018a). Normalized Cut Loss for Weakly-supervised CNN Segmentation. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1818–1827.

Tang, M., Perazzi, F., Djelouah, A., Ben Ayed, I., Schroers, C. & Boykov, Y. (2018b). On Regularized Losses for Weakly-supervised CNN Segmentation. *European Conference on Computer Vision (ECCV)*, 11220, 524–540.

Tasdizen, T., Sajjadi, M., Javanmardi, M. & Ramesh, N. (2018). Improving the robustness of convolutional networks to appearance variability in biomedical images. *International Symposium on Biomedical Imaging (ISBI)*.

Tong, J., Zhao, Y., Zhang, P., Chen, L. & Jiang, L. (2019). MRI brain tumor segmentation based on texture features and kernel sparse coding. *Biomedical Signal Processing and Control*, 47, 387–392.

Tsai, Y.-H. et al. (2018). Learning to adapt structured output space for semantic segmentation. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7472–7481.

Tzeng, E., Hoffman, J., Zhang, N., Saenko, K. & Darrell, T. (2014). Deep Domain Confusion: Maximizing for Domain Invariance. ArXiv 1412.3474.

Tzeng, E., Hoffman, J., Darrell, T. & Saenko, K. (2015). Simultaneous deep transfer across domains and tasks. *International Conference on Computer Vision (ICCV)*.

Tzeng, E., Hoffman, J., Saenko, K. & Darrell, T. (2017). Adversarial Discriminative Domain Adaptation. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2962–2971.

Vakalopoulou, M., Chassagnon, G., Bus, N., Marini Silva, R., Zacharaki, E. I., Revel, M.-P. et al. (2018). AtlasNet: Multi-atlas Non-linear Deep Networks for Medical Image Segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 658–666.

Van Tulder, G. & de Bruijne, M. (2016). Representation Learning for Cross-Modality Classification. *MICCAI Workshop on Medical Computer Vision*.

Varoquaux, G. & Cheplygina, V. (2022). Machine learning for medical imaging: methodological failures and recommendations for the future. *NPJ Digital Medicine*, 5(1), 48.

Varsavsky, T., Orbes-Arteaga, M., Sudre, C., Graham, M., Nachev, P. & Cardoso, M. (2020). Test-time unsupervised domain adaptation. *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 428–436.

Veksler, O. (2008). Star Shape Prior for Graph-Cut Image Segmentation. *European Conference on Computer Vision (ECCV)*, pp. 454–467.

Vlontzos, A., Rueckert, D. & Kainz, B. (2022). A Review of Causality for Learning Algorithms in Medical Image Analysis. *Machine Learning for Biomedical Imaging (MELBA)*, 1, 1–17.

Volpi, R., Namkoong, H., Sener, O., Duchi, J. C., Murino, V. & Savarese, S. (2018). Generalizing to unseen domains via adversarial data augmentation. *Advances in Neural Information Processing Systems (NeurIPS)*, 31, 5339–5349.

Volpi, R., de Jorge, P., Larlus, D. & Csurka, G. (2022). On the Road to Online Adaptation for Semantic Image Segmentation. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19184–19195.

Vu, T.-H., Jain, H., Bucher, M., Cord, M. & Pérez, P. (2019). ADVENT: Adversarial Entropy Minimization for Domain Adaptation in Semantic Segmentation. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2517–2526). IEEE.

Wachinger, C., Rieckmann, A. & Pölsterl, S. (2021). Detect and correct bias in multi-site neuroimaging datasets. *Medical Image Analysis*, 67, 101879.

Wachinger, C. et al. (2016). Domain adaptation for Alzheimer's disease diagnostics. *NeuroImage*, 139, 470–479.

Wagner, S. K., Hughes, F., Cortina-Borja, M., Pontikos, N., Struyven, R., Liu, X. et al. (2022). AlzEye: longitudinal record-level linkage of ophthalmic imaging and hospital admissions of 353 157 patients in London, UK. *BMJ Open*, 12(3), 1–1.

Wang, D., Shelhamer, E., Liu, S., Olshausen, B. & Darrell, T. (2021). Tent: Fully Test-Time Adaptation by Entropy Minimization. *International Conference on Learning Representations (ICLR)*.

Wang, Q., Fink, O., Van Gool, L. & Dai, D. (2022). Continual Test-Time Domain Adaptation. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7201–7211.

Wei, Y., Liang, X., Chen, Y., Shen, X., Cheng, M.-M., Zhao, Y. et al. (2016). STC: A Simple to Complex Framework for Weakly-Supervised Semantic Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 39, 2314–2320.

Winkler, J. K., Fink, C., Toberer, F., Enk, A., Deinlein, T., Hofmann-Wellenhof, R. et al. (2019). Association Between Surgical Skin Markings in Dermoscopic Images and Diagnostic Performance of a Deep Learning Convolutional Neural Network for Melanoma Recognition. *JAMA Dermatology*, 155(10), 1135.

Wu, K., Du, B., Luo, M., Wen, H., Shen, Y. & Feng, J. (2019). Weakly supervised brain lesion segmentation via attentional representation learning. *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 211–219.

Wu, X., Zhang, S., Zhou, Q., Yang, Z., Zhao, C. & Latecki, L. J. (2020). Entropy Minimization vs. Diversity Maximization for Domain Adaptation. arXiv 2002.01690.

Xiao, W., Huang, X., Wang, J. H., Lin, D. R., Zhu, Y., Chen, C. et al. (2021). Screening and identifying hepatobiliary diseases through deep learning using ocular images: a prospective, multicentre study. *The Lancet Digital Health*, 3(2), e88–e97.

Xie, Q., Luong, M.-T., Hovy, E. & Le, Q. V. (2020). Self-training with noisy student improves imagenet classification. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10687–10698.

Xie, S., Girshick, R., Dollar, P., Tu, Z. & He, K. (2017). Aggregated Residual Transformations for Deep Neural Networks. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1492–1500.

Xu, J., Xiao, L. & López, A. M. (2019). Self-supervised domain adaptation for computer vision tasks. *IEEE Access*, 7, 156694–156706.

Yan, W., Wang, Y., Gu, S., Huang, L., Yan, F., Xia, L. et al. (2019). The domain shift problem of medical image segmentation and vendor-adaptation by Unet-GAN. *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 623–631.

Yao, Y., Liu, F., Zhou, Z., Wang, Y., Shen, W., Yuille, A. et al. (2022). Unsupervised Domain Adaptation through Shape Modeling for Medical Image Segmentation. arXiv preprint 2207.02529.

Yap, J., Yolland, W. & Tschandl, P. (2018). Multimodal skin lesion classification using deep learning. *Experimental dermatology*, 27(11), 1261–1267.

Yi, X., Walia, E. & Babyn, P. (2019). Generative adversarial network in medical imaging: A review. *Medical Image Analysis*, 58, 101552.

Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J. & Oermann, E. K. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine*, 15(11), e1002683.

Zeng, Q., Karimi, D., Pang, E. H., Mohammed, S., Schneider, C., Honarvar, M. et al. (2019). Liver segmentation in magnetic resonance imaging via mean shape fitting with fully convolutional neural networks. *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 246–254.

Zhang, K., Schölkopf, B., Muandet, K. & Wang, Z. (2013). Domain adaptation under target and conditional shift. *International Conference on Machine Learning (ICML)*, pp. 819–827.

Zhang, W., Deng, L., Zhang, L. & Wu, D. (2022). A survey on negative transfer. *IEEE/CAA Journal of Automatica Sinica*, 10(2), 305–329.

Zhang, Y., Qiu, Z., Yao, T., Liu, D. & Mei, T. (2018a). Fully Convolutional Adaptation Networks for Semantic Segmentation. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6810–6818.

Zhang, Y., David, P., Foroosh, H. & Gong, B. (2020a). A Curriculum Domain Adaptation Approach to the Semantic Segmentation of Urban Scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 42, 1823–1841.

Zhang, Y., David, P. & Gong, B. (2020b). A curriculum domain adaptation for semantic segmentation of urban scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 42(8), 1823–1841.

Zhang, Y. et al. (2018b). Task driven generative modeling for unsupervised domain adaptation: Application to x-ray image segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 599–607.

Zhao, H. et al. (2019). Supervised Segmentation of Un-Annotated Retinal Fundus Images by Synthesis. *IEEE Transactions on Medical Imaging (TMI)*, 38(1), 46–56.

Zhao, S., Yue, X., Zhang, S., Li, B., Zhao, H., Wu, B. et al. (2020). A review of single-source deep unsupervised visual domain adaptation. *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*, 33(2), 473–493.

Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D. et al. (2015). Conditional random fields as recurrent neural networks. *International Conference on Computer Vision (ICCV)*, pp. 1529–1537.

Zhou, X.-Y., Guo, Y., Shen, M. & Yang, G.-Z. (2020). Application of artificial intelligence in surgery. *Frontiers of Medicine*, 14(4), 417–430.

Zhou, Y., Li, Z., Bai, S., Chen, X., Han, M., Wang, C. et al. (2019a). Prior-Aware Neural Network for Partially-Supervised Multi-Organ Segmentation. *International Conference on Computer Vision (ICCV)*, pp. 10672–10681.

Zhou, Y., Wang, Y., Tang, P., Bai, S., Shen, W., Fishman, E. et al. (2019b). Semi-Supervised 3D Abdominal Multi-Organ Segmentation Via Deep Multi-Planar Co-Training. *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 121–140.

Zhu, J., Park, T., Isola, P. & Efros, A. A. (2017). Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. *International Conference on Computer Vision (ICCV)*, pp. 2242–2251.

Zhuang, X. et al. (2019). Evaluation of algorithms for Multi-Modality Whole Heart Segmentation: An open-access grand challenge. *Medical Image Analysis*, 58, 101537.

Zou, Y., Yu, Z., Kumar, B. V. K. V. & Wang, J. (2018). Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. *European Conference on Computer Vision (ECCV)*, pp. 289–305.