Medical Image Segmentation Under Challenging Scenarios

by

Ping WANG

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

MONTREAL, SEPTEMBER 28, 2023

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

BOARD OF EXAMINERS

THIS THESIS HAS BEEN EVALUATED

BY THE FOLLOWING BOARD OF EXAMINERS

Mr. Christian Desrosiers, Thesis supervisor
Department of Software and IT Engineering, École de technologie supérieure

Mr. Marco Pedersoli, Thesis Co-Supervisor
Department of Systems Engineering, École de technologie supérieure

Mr. Caiming Zhang, Thesis Co-Supervisor
Department of Software, Shandong University

Mr. Eric Granger, Chair, Board of Examiners
Department of Systems Engineering, École de technologie supérieure

Mr. Jose Dolz, Member of the Jury
Department of Software and IT Engineering, École de technologie supérieure

Mr. Nicolas Thome, External Examiner
Department of Computer Science, Sorbonne University

THIS THESIS WAS PRESENTED AND DEFENDED

IN THE PRESENCE OF A BOARD OF EXAMINERS AND THE PUBLIC

ON SEPTEMBER 15, 2023

AT ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

ACKNOWLEDGEMENTS

I would like to express my deepest appreciation to my supervisors, Prof.Christian Desrosiers and Prof. Marco Pedersoli, for their invaluable guidance, unwavering support, and expert mentorship throughout the course of this research. Their insightful feedback and constructive suggestions have been great in shaping the direction and quality of this work. Owing to Prof. Desrosiers, I was able to come to LIVIA for pursuing my PhD degree. Thanks to his help, I obtained a competitive scholarship without which it would have been hard to focus entirely on research. I also appreciate the patience and encouragements of Prof. Desrosiers when I started my studies in Canada, trying to improve my English speaking skills. Prof. Pedersoli also guided my research by offering brilliant insights and with his attention to detail. I am particularly grateful for the open questions he posed, encouraging me to have a deeper contemplation and expanding my thinking.

I would also like to thank my co-supervisors, Prof. Caiming Zhang and Prof. Yunfeng Zhou, for giving me the chance to have the internship in the Intelligent Graphics and Image Processing Lab of the Shandong University. I also appreciate the help they offered when I needed access to the University during COVID.

I would also like to thank members of the jury for evaluating my thesis and supporting my defense. I also thank Prof. Jose Dolz and Prof. Eric Granger for the valuable and constructive suggestions during the presentation of my PhD proposal.

I further wish to thank Jizong Peng, a PhD supervised by Prof. Desrosiers and Prof. Pedersoli for his great help both in research and life. I am grateful for his help to learn Python programming and PyTorch, as well as for his guidance on organizing experiments. I also appreciate his constructive advises on research methodology and the help he offered to me in life when I first came to ÉTS.

Again, I would like to thank the ÉTS for offering me financial support from Development of International Joint Research Projects throughout my studies. I also thank Compute Canada and

the Department of Software and IT engineering of ÉTS for granting me large-scale computation resources. Most of my works would not have been accomplished without this support.

Finally, I am profoundly grateful to my family. Their unconditional love, endless encouragement, and sacrifices have been the foundation of my personal and academic achievements. I am forever grateful for their belief in me and for always being there, providing unwavering support and understanding.

Segmentation des images médicales dans des scénarios difficiles

Ping WANG

RÉSUMÉ

La segmentation est une tâche critique dans l'analyse d'images médicales, qui joue un rôle essentiel dans le diagnostic assisté par ordinateur, la radiothérapie guidée par l'image et la navigation chirurgicale. Les méthodes de segmentation basées sur l'apprentissage profond ont réalisé des progrès sans précédents, bénéficiant d'une grande quantité de données annotées. Or, l'acquisition de données annotées en imagerie médicale requiert des efforts et coûts substantiels, limitant l'application de l'apprentissage profond pour la segmentation d'images dans ce domaine. Pour obtenir des performances compétitives avec des données annotées limitées, des approches à base d'apprentissage semi-supervisé ont été développées pour exploiter également les données non annotées, Bien que ces approches aient permis d'améliorer les performances, l'entraînement non supervisé sur des données sans annotation apporte certains défis. Par exemple, des prédictions incorrectes sur des données non annotées, lors de la phase d'entraînement initiale, peuvent être accentuées au fur et à mesure que l'entraînement progresse, provoquant une dégradation des performances. Un autre problème est que les données annotées peuvent être insuffisantes pour que le modèle apprenne une forme anatomiquement plausible de l'organe à segmenter. Dans ce cas, il peut être utile d'employer des apriori anatomiques pour guider l'apprentissage du modèle. En outre, une autre limitation pratique est que les données annotées et non annotées peuvent avoir des distributions distinctes en raison de différences dans les appareils d'acquisition d'images. Cela représente une tâche plus complexe car l'apprentissage du modèle peut être dominé pour les données annotées et, conséquemment, ce dernier ne pourra pas s'adapter efficacement à la distribution des données à segmenter. Plusieurs approches ont été proposées pour l'adaptation de domaine, par exemple, utilisant des décodeurs de reconstruction auxiliaires ou des techniques à base de transfert de style. Cependant, l'exploration de solutions simples mais efficaces pour ce problème, qui évitent l'emploi d'une infrastructure logicielle complexe, demeure une direction critique de recherche.

L'objectif principal de cette thèse est de développer des méthodes simples et précises pour la segmentation d'images médicales dans ces deux scénarios difficiles. Dans ce but, nous proposons tout d'abord une méthode de co-entraînement auto-rythmée et auto-cohérente pour la segmentation semi-supervisée d'images. Cette méthode résout le problème des prédictions incorrectes pour les données non étiquetées au cours de la phase d'entraînement initiale, améliorant ainsi la segmentation. Deuxièmement, nous avons développé une approche d'entraînement antagoniste avec contraintes pour la segmentation semi-supervisée anatomiquement plausible. Cette approche permet d'obtenir une segmentation anatomiquement plausible en incorporant des apiori anatomiques complexes non différentiables. La dernière contribution se concentre sur le scénario plus difficile de l'adaptation de domaine. Pour cette tâche, nous avons proposé une méthode d'alignement de distribution conjointe sensible à la forme pour la segmentation inter-domaines d'images. Cette méthode atteint des performances compétitives de segmentation inter-domaines en alignant explicitement la représentation invariante de domaine modélisant la

taille des classes de segmentation et la relation spatiale entre ces classes. Cette thèse a donné lieu à trois publications dans des revues de haut niveau en imagerie médicale et une publication à une des principales conférences dans ce domaine. Les objectifs spécifiques de cette thèse sont présentés ci-dessous.

Dans notre premier objectif, nous nous concentrons sur la segmentation semi-supervisée et proposons une méthode basée sur un cadre de co-entraînement. Tout d'abord, une stratégie d'apprentissage auto-rythmée pour le co-entraînement est présentée, permettant aux réseaux de neurones entraînés conjointement de se concentrer d'abord sur les régions les plus faciles à segmenter, puis de tenir compte progressivement des régions plus difficiles. Ceci est mis en œuvre via une fonction de perte différentiable de bout en bout sous la forme d'une divergence Jensen Shannon généralisée (JSD). Pour encourager les réseaux à produire non seulement des prédictions cohérentes mais aussi ayant une haute confiance, nous améliorons cette perte JSD généralisée avec un régularisateur d'incertitude basé sur l'entropie. La robustesse des modèles individuels dans le cadre de co-entraînement est ensuite améliorée à l'aide d'une stratégie par ensemble temporel qui force leur prédiction à être cohérente à travers différentes itérations de l'entraînement. L'efficacité de cette méthode est évaluée sur trois jeux de données de segmentation complexe comprenant des images de différentes modalités, pour lesquels elle améliore la précision de la segmentation lorsque très peu d'images annotées sont utilisées. Nous explorons également l'impact de la stratégie d'apprentissage auto-rythmée proposée, de la stratégie d'auto-cohérence, ainsi que du régularisateur d'incertitude proposé. Les résultats expérimentaux montrent l'efficacité de chacun des composants de la méthode proposée.

Notre deuxième objectif porte également sur la segmentation semi-supervisée. Pour cet objectif, une méthode d'entraînement antagoniste avec contraintes est proposée pour la segmentation semi-supervisée anatomiquement plausible. Contrairement aux approches se concentrant uniquement sur les mesures de précision comme le Dice, cette méthode prend en compte des contraintes anatomiques complexes telles que la connectivité, la convexité et la symétrie qui ne peuvent pas être facilement modélisées dans une fonction de perte. Le problème des contraintes non différentiables est résolu à l'aide d'un algorithme de renforcement qui permet d'obtenir un gradient pour les contraintes violées. Pour générer de manière dynamique des exemples violant les contraintes, et ainsi obtenir des gradients utiles à l'apprentissage, notre méthode adopte une stratégie d'entraînement antagoniste qui modifie les images d'entraînement pour maximiser la perte de contrainte, puis met à jour le réseau pour qu'il soit robuste à ces exemples antagonistes. La méthode proposée offre un moyen générique et efficace d'ajouter des contraintes de segmentation complexes par dessus n'importe quel réseau de segmentation. Des expériences sur des données synthétiques et quatre ensembles de données cliniquement pertinentes démontrent l'efficacité de notre méthode en termes de précision de segmentation et de plausibilité anatomique.

Le dernier objectif se concentre sur le scénario d'adaptation de domaine. Pour ce scénario, nous avons développé une méthode d'alignement de distribution conjointe sensible aux formes pour la segmentation inter-domaines. Cette méthode aligne les statistiques d'ordre élevé, calculées pour les domaines source et cible, qui encodent les relations spatiales invariantes au domaine

entre les classes de segmentation. Notre méthode estime d'abord la distribution conjointe des prédictions pour une paire de pixels dont la position relative correspond à un déplacement spatial donné. L'adaptation de domaine est alors réalisée en alignant les distributions conjointes des images source et cible, calculées pour un ensemble de déplacements. Deux améliorations de cette méthode sont ensuite proposées. La première utilise une stratégie multi-échelle efficace qui permet de capturer les relations à longue portée dans les statistiques. La seconde étend la perte d'alignement de distribution conjointe aux caractéristiques dans les couches intermédiaires du réseau, en calculant leur intercorrélation. Nous testons notre méthode sur la tâche de segmentation cardiaque multimodale non appariée à l'aide de l'ensemble de données Multi-Modality Whole Heart Segmentation Challenge (MMWHS) et sur la tâche de segmentation de la prostate, où les images de deux ensembles de données sont employées comme données provenant de différents domaines. Nos résultats montrent les avantages de notre méthode par rapport aux approches récentes de segmentation inter-domaines d'images.

Mots-clés: segmentation des images médicales, segmentation semi-supervisée, adaptation de domaine non supervisée, contraintes non différentiables, représentation invariante de domaine

Medical Image Segmentation Under Challenging Scenarios

Ping WANG

ABSTRACT

Segmentation is a critical task in medical image analysis, which plays a vital role in computer-aided diagnose, image-guided radiotherapy, and surgical navigation. Deep learning-based segmentation methods have achieved unprecedented progress in recent years, benefiting from large amounts of annotated data. However, obtaining annotations for medical images requires substantial efforts and costs. Further, the limited availability of annotated medical data poses a significant challenge in achieving high-performance medical image segmentation. To obtain competitive performance with limited labeled data, semi-supervised learning approaches have been developed to also exploit unlabeled data. Though these approaches have achieved an improved performance, the unsupervised training on unlabeled data also brought some challenges. For instance, inaccurate predictions made for unlabeled data in the initial training stage can be accentuated as the training progresses, leading to a degradation in performance. Another problem is the labeled data may be insufficient for the model to learn an anatomical-plausible shape for the organ to segment. In such case, it may be useful to employ anatomical priors to guide the model learning. Another practical limitation is that the labeled data and unlabeled data can have distinct distributions due to differences in the image acquisition devices. This represents a more challenging task since the model's training can be dominated by labeled data and, as a result consequently, this model may fail to adapt to the distribution of target data. Several approaches have been proposed for domain adaptation, for instance, relying on auxiliary reconstruction decoders or style-transfer. However, developing simple yet highly effective solutions for this task, that avoid the use of a complex framework, is still a pressing direction of research.

The main objective of this thesis is to develop simple and accurate methods for medical image segmentation under these two challenging scenarios. Specifically, we first proposed a self-paced and self-consistent co-training method for semi-supervised image segmentation. This method addresses the problem of inaccurate predictions for unlabeled data during the initial training stage, thereby boosting segmentation performance. Secondly, we developed a constrained adversarial training method for semi-supervised segmentation, which enforces anatomical-plausible predictions by incorporating complex non-differentiable anatomical priors. The last contribution focuses on the more challenging domain adaptation scenario. For this task, we proposed a shape-aware joint distribution alignment method for cross-domain image segmentation, which achieves competitive cross-domain segmentation performance by explicitly aligning domain-invariant representation encoding shape size and spatial relationship between classes. This thesis has resulted in three publications in high-impact medical imaging journals as well as a publication in a top conference of that field. The specific objectives of this thesis are presented below.

In our first objective, we focus on semi-supervised segmentation and propose a method based on a co-training framework. First, we present a self-paced learning strategy for co-training that enables jointly-trained neural networks to focus on easier-to-segment regions first, and then gradually consider harder ones. This strategy is implemented via an end-to-end differentiable loss in the form of a generalized Jensen Shannon Divergence (JSD). To encourage the networks to produce not only consistent but also confident predictions, we enhance this generalized JSD loss with an uncertainty regularizer based on entropy. Furthermore, the robustness of individual models in our co-training framework is further improved using a self-ensembling loss that enforces the models prediction to be consistent across different training iterations. The effectiveness of our method is assessed on three challenging segmentation datasets including images of different modalities, for which it boosts segmentation accuracy when very few labeled images are used. We also explore the impact of the proposed self-paced learning strategy, self-consistency strategy, as well as our uncertainty regularizer. Experimental results show the effectiveness of each component in the proposed method.

Our second objective also focuses on semi-supervised segmentation. For this objective, a constrained adversarial training method is proposed for anatomical-plausible segmentation. Unlike approaches focusing solely on accuracy measures like Dice, this method considers complex anatomical constraints like connectivity, convexity, and symmetry that cannot be easily modeled in a loss function. The problem of non-differentiable constraints is solved using the REINFORCE algorithm which enables to obtain a gradient for the violated constraints. To generate constraint-violating examples on the fly, and thus obtain useful gradients, our method adopts an adversarial training strategy which modifies training images to maximize the constraint loss, and then updates the network to be robust to these adversarial examples. The proposed method offers a generic and efficient way to add complex segmentation constraints on top of any segmentation network. Experiments on four clinically-relevant datasets as well as on synthetic datasets generated for this work demonstrate the effectiveness of our method in terms of segmentation accuracy and anatomical plausibility.

The last objective focuses on the domain adaptation scenario. For this scenario, we developed a shape-aware joint distribution alignment method for cross-domain segmentation, which aligns high-order statistics, computed for the source and target domains, that encode domain-invariant spatial relationships between segmentation classes. Our method first estimates the joint distribution of predictions for pairs of pixels whose relative position corresponds to a given spatial displacement. Domain adaptation is then achieved by aligning the joint distributions of source and target images, computed for a set of displacements. Two enhancements of this method are proposed. The first one uses an efficient multi-scale strategy that enables capturing long-range relationships in the statistics. The second one extends the joint distribution alignment loss to features in intermediate layers of the network by computing their cross-correlation. We test our method on the task of unpaired multi-modal cardiac segmentation using the Multi-Modality Whole Heart Segmentation (MMWHS) Challenge dataset and on the task of prostate segmentation task, where images of two datasets are taken as data from different domains. Our results show the advantages of our method compared to recent approaches for cross-domain image segmentation.

Keywords: medical image segmentation, semi-supervised segmentation, unsupervised domain adaptation, non-differentiable constraints, domain-invariant representation

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

LIST OF ALGORITHMS

Page

LIST OF ABBREVIATIONS

ASSD	Average Symmetric Surface Distance
CAT	Constrained Adversarial Training
CT	Computed Tomography
CNNs	Convolutional Neural Networks
DSC	Dice Similarity Coefficient
FCN	Fully Convolutional Network
GAN	Generative Adversarial Network
GT	Ground Truth
JSD	Jensen-Shannon Divergence
KL	Kullback-Leibler
LDS	Local Distribution Smoothness
MHD	Modified Hausdorff Distance
MRI	Magnetic Resonance Imaging
MSE	Mean Squared Error
MMD	Maximum Mean Discrepancy
SGD	Stochastic Gradient Descent
SPL	Self-paced Learning
UDA	Unsupervised Domain Adaptation
VAE	Variational Auto-Encoder
VAT	Virtual Adversarial Training

LIST OF SYMBOLS AND UNITS OF MEASUREMENTS

$\mathcal{D}$	Training dataset
$\mathcal{S}$	Labeled or source dataset
$\mathcal{U}$	Unlabeled dataset
$\mathcal{T}$	Target dataset
θ	Model's parameters
Ω	Pixels
C	Semantic classes
$\mathbf{y}$	Ground truth
$\widehat{\mathbf{y}}$	Prediction
$\mathcal{B}$	Training batch
$\mathcal{L}$	Loss
$\mathcal{H}$	Entropy

INTRODUCTION

0.1 Semantic segmentation for medical images

Semantic segmentation (Winn & Shotton, 2006; Long, Shelhamer & Darrell, 2015), a computer vision task, involves dividing an image into meaningful and distinct regions or segments where each pixel in the image is assigned a label corresponding to a specific object or class. The goal of this task is to understand the semantic meaning and structure of the scene by classifying and segmenting different objects or regions within the image. For medical images, semantic segmentation can help delineate organs or lesions, as shown in Figure 0.1. It plays an essential role in various applications of medical image analysis, such as organ segmentation, tumor or lesion detection and segmentation, image-guided radiotherapy, and surgical navigation.

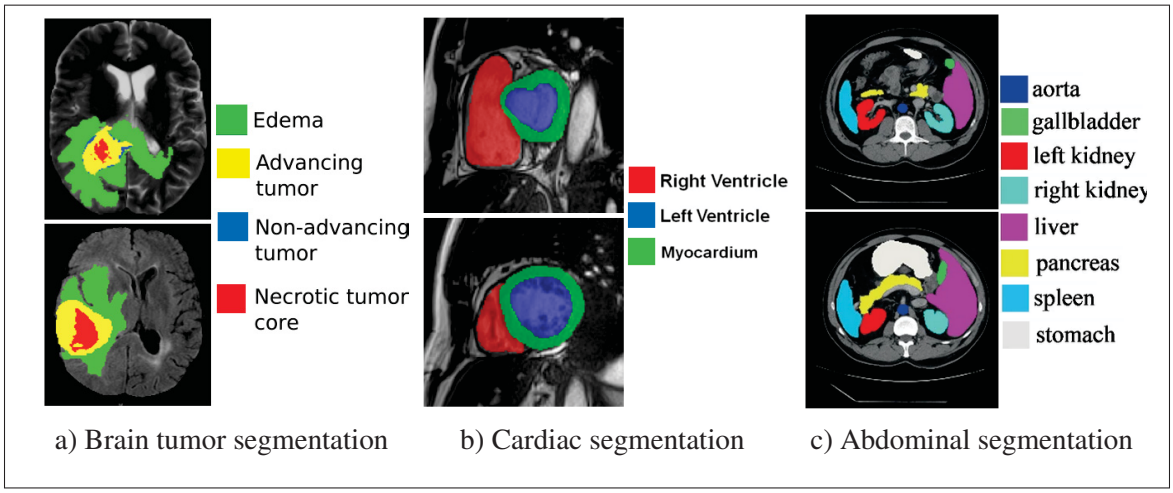


Figure 0.1 Visualization of semantic segmentation for medical image.

Among the various methods developed for medical image segmentation, those based on deep learning (Goodfellow, Bengio & Courville, 2016) have obtained the most success. Deep learning-based medical image segmentation methods typically leverage convolutional neural networks (CNNs) (LeCun, Bottou, Bengio & Haffner, 1998) or vision transformers (ViT) (Hatamizadeh *et al.*, 2022) to automatically and accurately segment various structures or abnormalities. Usually,

the CNN or ViT network takes a raw image as input and outputs the corresponding dense prediction (e.g. labels for each pixel). The knowledge to perform a segmentation task is learned from data into the network parameters, by minimizing a loss function that forces the dense prediction to be as close as possible to the ground-truth labels (Rumelhart, Hinton & Williams, 1986).

0.2 Challenges and problem statement

With the use of deep learning techniques, medical image segmentation has made remarkable advancements, attaining outstanding performance and accuracy. However, when confronted with real-world scenarios of medical image segmentation, several challenges arise. First, CNNs require a large amount of dense ground-truth annotations during training to properly learn the task. However, unlike image-level labeling for classification tasks, the pixel-wise annotation of data for semantic segmentation can be costly and time-consuming. Furthermore, annotations for medical data are highly dependent on expert guidance. As a result, real-world scenarios often involve a limited availability of labeled data. However, a substantial amount of unlabeled data often remains at hand, which can be exploited in the semi-supervised setting (Rasmus, Berglund, Honkala, Valpola & Raiko, 2015). Naturally, the first challenge is how to obtain a high accuracy segmentation in this scenario.

Another challenge comes from the variance observed in medical imaging. Medical images can vary widely in terms of acquisition protocols, imaging devices, patient populations, and pathological conditions. As an example, medical imaging modalities, such as computed tomography (CT), magnetic resonance imaging (MRI), ultrasound, and X-ray exhibit distinctive statistical properties. The presence of differences across data distributions, also called domain shift (Ben-David *et al.*, 2010), can affect the performance of segmentation models trained on a given dataset when applied to another dataset, resulting in a significant decrease in segmentation accuracy.

In light of these challenges, this thesis focuses on developing segmentation methods for medical images that can achieve both high accuracy and anatomically-plausible segmentation in scenarios with limited annotations, and obtaining state-of-art performance for cross-domain segmentation tasks.

0.3 Motivations and objectives

As highlighted previously, the objective of this thesis is to develop segmentation methods under the aforementioned challenges. We achieve this objective by decomposing it into three specific objectives. The first objective aims to develop a semi-supervised segmentation method based on self-paced learning and self-consistent co-training, generating state-of-the-art performance for scenarios where there is a little amount of labeled data and a large amount of unlabeled data. The second objective aims to avoid anatomically-impossible segmentation under the semi-supervised scenario by leveraging anatomically-complex constraints and virtual adversarial training. The third objective is to propose a cross-domain segmentation method by aligning domain-invariant statistics encoding shape size and spatial relationships, modeled by a joint probability distribution and cross-correlation. These objectives are detailed as follows.

Objective 1: Our first objective is to present an efficient semi-supervised method which can improve semantic segmentation when only a few labeled images are available for training. Co-training (Blum & Mitchell, 1998) is a popular method for semi-supervised learning, which encourages consistent predictions for two independent views of the data. Existing semi-supervised segmentation methods based on co-training improve the segmentation performance based on two main strategies: increasing the diversity of different views with supports from adversarial examples (Peng, Estrada, Pedersoli & Desrosiers, 2020a; Xie *et al.*, 2023) and encouraging confident prediction consistency with the aid of uncertainty-aware mechanisms (Zheng *et al.*, 2022; Xia *et al.*, 2020a). However, current co-training methods do not employ a self-paced learning strategy in which easier regions first and gradually-harder regions are involved during

training, and thus are susceptible to incorrect predictions during the initial training stage. Moreover, these methods often overlook the robustness of each view, and do not exploit the self-consistency within each individual model. In order to address these limitations, we aim to propose a self-paced and self-consistent co-training method for semi-supervised medical image segmentation. Our method focuses on gradually-harder regions of unlabeled data and encourages both consistency and confidence across co-trained models during training. By dynamically controlling the importance of individual pixels in the co-training of separate models forming an ensemble, our method can boost the performance for each model and yield a performance close to the fully-supervised setting.

Objective 2: Our second objective is to present a semi-supervised segmentation method that avoids anatomically-impossible predictions, while maintaining a competitive segmentation accuracy in terms of standard metrics. Medical images requires stricter segmentation results compared to natural images due to the critical nature of the analysis. Even though recent semi-supervised methods for medical image segmentation achieve a high performance (Antti & Valpola, 2017; Miyato, Maeda, Koyama & Ishii, 2019; Gao *et al.*, 2021), these methods may still generate predictions that are considered anatomically invalid by clinicians. For instance, a segmentation with high Dice similarity may still contain holes or disconnected regions that are anatomically impossible for an organ. In order to overcome this issue, we propose a constrained adversarial training method for semi-supervised segmentation, where complex, non-differentiable constraints representing anatomical priors are integrated into the segmentation using virtual adversarial training (VAT) (Miyato *et al.*, 2019) and the REINFORCE algorithm (Williams, 1992). Our method can add these complex constraints on top of any segmentation network, trained with standard back-propagation, to generate anatomically-plausible segmentations.

Objective 3: For our third objective, which focuses on domain adaptation (Ben-David *et al.*, 2010), we present an unsupervised domain adaptation (UDA) method that can achieve cross-

domain segmentation by aligning the domain-invariant representation between the source and target domains. Segmentation models trained on source domain data typically suffer from a severe decrease in performance when applied to data from the target domain due to the problem of domain shift (i.e, distribution differences between source and target domains). However, despite the presence of a domain shift, the anatomical structures generated for the same tissue or organ but from different modalities should remain consistent. Based on this idea, a typical method for domain adaptation seeks to align a domain-invariant representation based on disentangled representation learning (Yang *et al.*, 2019a; Dai *et al.*, 2021a). This alignment can be achieved implicitly with the help of an auxiliary reconstruction decoder or via a style transfer strategy. Instead of relying on a relatively complex technique for the implicit alignment of domain-invariant representations, a recent method exploits class-level alignment (Bateson, Kervadec, Dolz, Lombaert & Ben Ayed, 2020) in a simpler, explicit manner. While it imposes the classes of the source and target domains to have the same relative sizes, this method overlooks the spatial relationships between these classes. To address this limitation, we propose a shape-aware joint distribution alignment approach for cross-domain image segmentation. Our approach explicitly aligns domain-invariant representations encoding both the size and spatial relationships of segmentation classes via a joint probability distribution. Using this simple, yet powerful strategy, our approach achieves state-of-the-art performance for the cross-domain segmentation of medical images.

0.4 Publications

The research presented in this thesis has led to the publication of three first-authored papers in high-impact medical imaging journals, i.e., Medical Image Analysis (MIA) and IEEE Transaction on Medical Imaging (TMI), and a paper in a leading conference of this field, i.e., International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI).

Moreover, an extension of this work was published in the highest-ranking conference in machine learning, i.e., Neural Information Processing Systems (NeurIPS).

Journal papers

- P. Wang, J. Peng, M. Pedersoli, Y. Zhou, C. Zhang and C. Desrosiers, Self-paced and self-consistent co-training for semi-supervised image segmentation, Medical Image Analysis, vol. 73, 2021, doi.org/10.1016/j.media.2021.102146.
- P. Wang, J. Peng, M. Pedersoli, Y. Zhou, C. Zhang and C. Desrosiers, CAT: Constrained adversarial training for anatomically-plausible semi-supervised segmentation, IEEE Transactionson Medical Imaging, doi: 10.1109/TMI.2023.3243069.
- P. Wang, J. Peng, M. Pedersoli, Y. Zhou, C. Zhang and C. Desrosiers, Shape-aware joint distribution alignment for cross-domain image segmentation, IEEE Transactions on Medical Imaging, doi: 10.1109/TMI.2023.3247941.

Conference papers

- P. Wang, J. Peng, M. Pedersoli, Y. Zhou, C. Zhang and C. Desrosiers, Context-aware virtual adversarial training for anatomically-plausible segmentation, Medical Image Computing and Computer Assisted Intervention, 2021
- J. Peng, P. Wang, C. Desrosiers, and M. Pedersoli, Self-paced contrastive learning for semi-supervised medical image segmentation with meta-labels, Advances in Neural Information Processing Systems, 2021.

0.5 Outline

The rest of this thesis is divided in five chapters, followed by an appendix.

Chapter 1 presents useful concepts and background knowledge needed for understanding the context and challenges of problems addressed in the thesis.

Chapter 2 presents our first contribution, a co-training framework for semi-supervised segmentation implementing self-paced learning strategy allowing jointly-trained neural networks to focus on easier-to-segment regions first, and then gradually consider harder ones. This framework also incorporates a self-consistency based on temporal ensembling to help distillate information from unlabeled images and obtain state-of-the-art performance.

Chapter 3 presents our second contribution, a constrained adversarial training method for anatomically-plausible semi-supervised segmentation. This method considers complex anatomical constraints which cannot be easily modeled in a differentiable loss function, using a REINFORCE algorithm that enables the model to obtain a gradient for violated constraints. We adopt an adversarial training strategy to generate adversarial samples that violate constraints in training, encouraging model to produce anatomically plausible segmentations.

Chapter 4 presents our last contribution, a shape-aware joint distribution alignment method for cross-domain image segmentation, in which the model realizes cross domain segmentation by aligning high-order statistics that encode domain-invariant spatial relationships of segmentation classes. We also extend this joint distribution alignment loss to features space alignment by computing their cross-correlation. The complementary of these two alignment strategies allows our method to achieve state-of-art performance for the cross-domain segmentation of medical images.

Conclusion and recommendations discusses the main contributions and limitation of this work, and proposes some potential directions of research to extend it.

Appendix I provides the detailed proof of theorems in our first contribution.

CHAPTER 1

BACKGROUND

In this chapter, we present some useful concepts and background knowledge needed for understanding the context and challenges of problems addressed in the thesis, including basic concepts of deep learning, common network architectures for medical image segmentation, techniques for self-paced learning and learning from non-differential losses, as well as the semi-supervised and unsupervised domain adaptation scenarios for image segmentation.

1.1 Basic concepts of deep learning

The development of deep learning has been marked by significant milestones and breakthroughs, leading to its wide application across various tasks such as semantic segmentation. In this section, we present some basic concepts of deep learning, including Convolutional neural networks (CNNs), loss functions, and optimization algorithms.

1.1.1 Convolutional neural network

CNNs are a specialized type of deep neural networks designed for processing structured grid-like data, particularly images. They have been tremendously successful in computer vision tasks, such as image classification, object detection, and image segmentation. Here, we briefly introduce some key components of CNNs.

Convolution: The convolution is a fundamental mathematical operation that plays a vital role in extracting features from input data. In simple terms, convolution involves applying a sliding window (also called a kernel or filter) over the input data and performing a dot product between the values in the window and the corresponding sub-region of the input. This operation captures local patterns and relationships between neighboring elements. Figure 1.1 shows a convolution operation with one kernel.

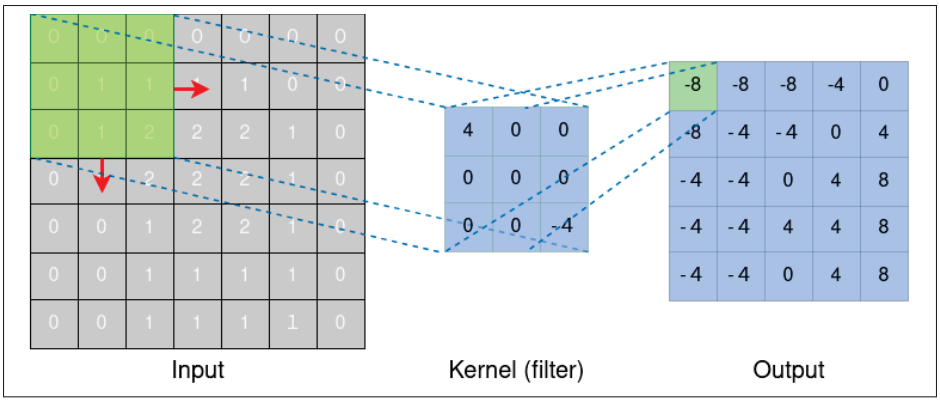


Figure 1.1 Convolution with a single kernel.

In CNNs, a convolutional layer always adopts multiple kernels, which enables this network type to learn a rich set of local patterns and feature representations. Figure 1.2 shows convolution operations with multiple kernels. To intuitively understand the representation ability of CNNs, we show kernels learned by the first convolutional layer of AlexNet on ImageNet (Krizhevsky, Sutskever & Hinton, 2017) in Figure 1.3. We can observe that a variety of frequency- and orientation-selective kernels, as well as various colored blobs are learned.

In summary, convolution is a key operation in CNNs as it helps the network effectively extract relevant features from the input data, allowing for hierarchical and spatially-informed representation learning.

Pooling: Pooling is a common component of CNNs used in various computer vision tasks. The primary purpose of pooling is to reduce the spatial dimensions of feature maps, while retaining important information. Two commonly used pooling operations are max pooling and average pooling, illustrated in Figure 1.4. As shown in the figure, the max pooling returns the maximum value in the region defined by a pooling size (2×2 in the example), and the average pooling returns the average value of the region, both reducing the dimensionality. Therefore, it enables faster computation and reduces the memory requirements of the network.

Normalization: Normalization is another commonly-used technique in CNNs, which normalizes the input data to a consistent scale or distribution, making it easier for the network to learn

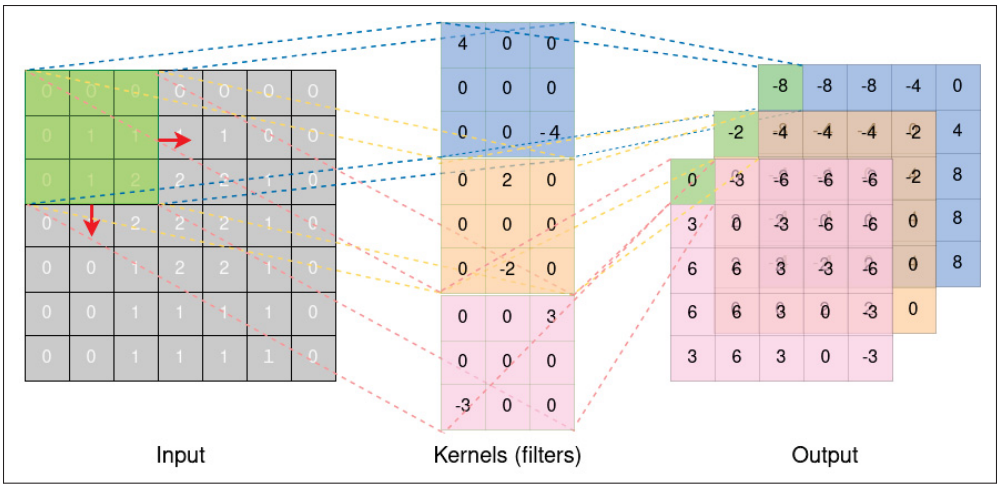


Figure 1.2 Convolution with multi-kernel (e.g. three kernels).

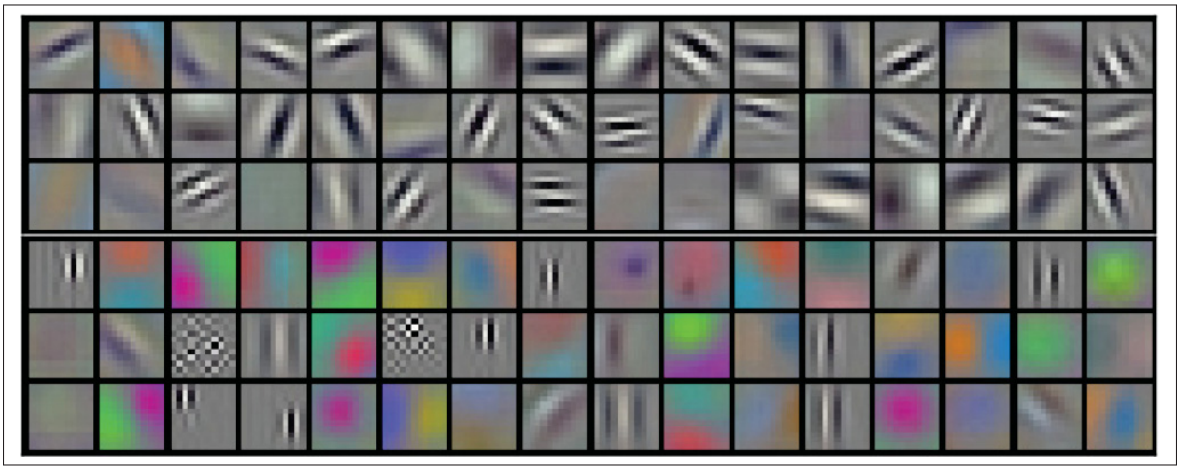


Figure 1.3 Visualization of kernels learned by the first convolutional layer of AlexNet on ImageNet. Image is taken from Krizhevsky *et al.* (2017).

and generalize patterns across different samples. There are four main types of normalization strategies in CNNs, batch normalization (Ioffe & Szegedy, 2015), layer normalization (Ba, Kiros & Hinton, 2016), instance normalization (Ulyanov, Vedaldi & Lempitsky, 2016), and group normalization (Wu & He, 2018), as shown in Figure 1.5. In batch normalization, which is often used under the assumption that samples within a batch are independent and identically distributed, the feature map tensor from a layer is normalized across the batch dimension by subtracting the batch mean and dividing by the batch standard deviation. For layer normalization,

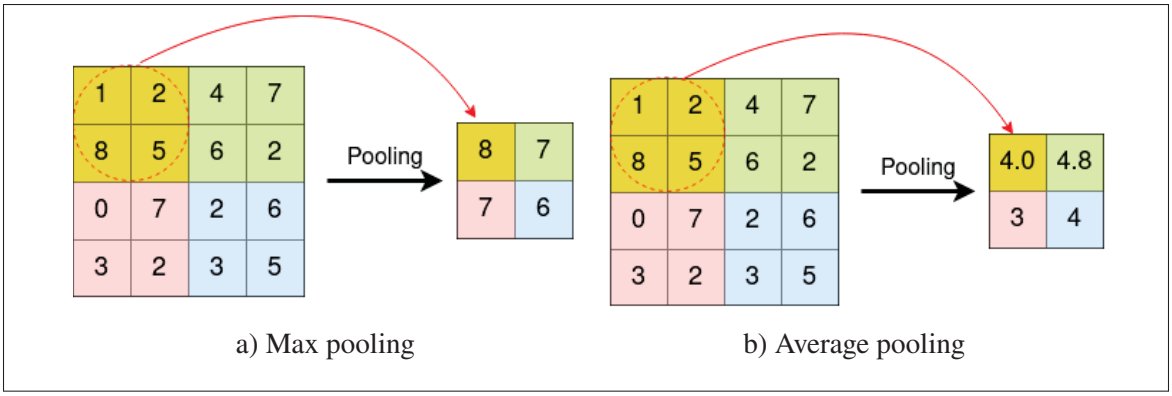


Figure 1.4 Max pooling and average pooling.

the feature map tensor is normalized within a layer by computing the mean and standard deviation across the spatial dimensions for each sample, whereas for instance normalization, the normalization is applied independently to each sample in a batch and each channel in the feature map. These two last normalization methods are particularly successful in training recurrent neural network (RNN) models or generative adversarial network (GAN) models (Ba *et al.*, 2016; Ulyanov *et al.*, 2016). Group normalization divides the channels into groups and performs normalization within each group. It aims to reduce the dependency on batch statistics, making it useful in scenarios where batch size is small.

Each normalization method has its own characteristics and applications. We can select different normalization methods in CNNs according to our datasets, tasks, network architectures, computational considerations, and so on.

Non-linear activation: The convolution is a linear operation, which does not allow learning and representing complex relationships between inputs and outputs. However, the relationship between the input and output very often exhibits non-linearity. In order to enable the network to learn non-linearity, non-linear activation is introduced into CNNs. Commonly used non-linear activation functions include sigmoid (McCulloch & Pitts, 1943), hyperbolic tangent (tanh), rectified linear unit (ReLU) (Nair & Hinton, 2010), softmax, and so on. The sigmoid function is

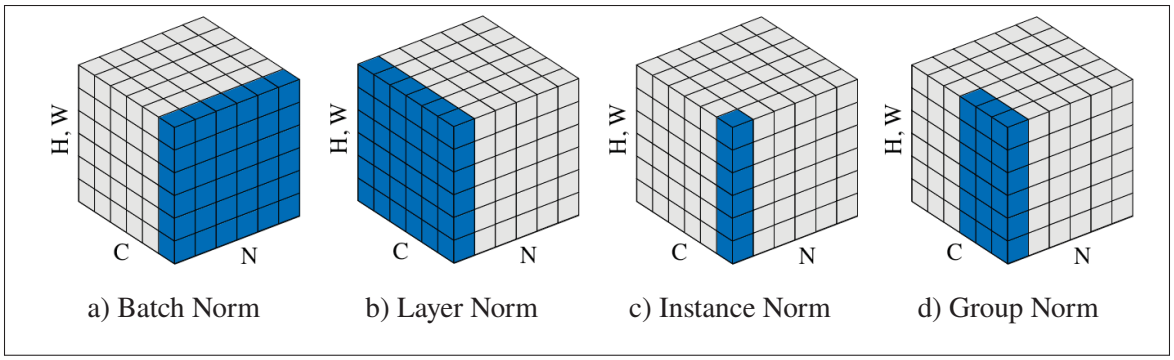


Figure 1.5 Normalization methods. Each subplot shows a feature map tensor, with N as the batch axis, C as the channel axis, and (H, W) as the spatial axes. The pixels in blue are normalized by the same mean and variance, computed by aggregating the values of these pixels. Image is taken from Wu & He (2018).

defined as

$$\sigma(x) = \frac{1}{1 + e^{-x}}. \tag{1.1}$$

The corresponding curve is shown in Figure 1.6 (a). We see that the sigmoid maps real-valued numbers to a range between 0 and 1. As x approaches positive infinity, the sigmoid function approaches 1, and as x approaches negative infinity, it approaches 0. Moreover, the derivative of the sigmoid function has a simple form expressed with itself $\sigma'(x) = \sigma(x)(1 - \sigma(x))$, which benefits optimization algorithms like gradient decent. However, a problem with the sigmoid is that it leads to vanishing gradients in deep networks, due to the near-zero gradient in the saturation regions.

The tanh function is defined as

$$\tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}. \tag{1.2}$$

Its curve is shown in Figure 1.6 (b). Different from the sigmoid, it maps real-valued numbers to a range between -1 and 1. Moreover, it has a steeper gradient around its midpoint comparing with sigmoid, allowing for more efficient learning and better gradient propagation in neural networks.

The ReLU function is defined as

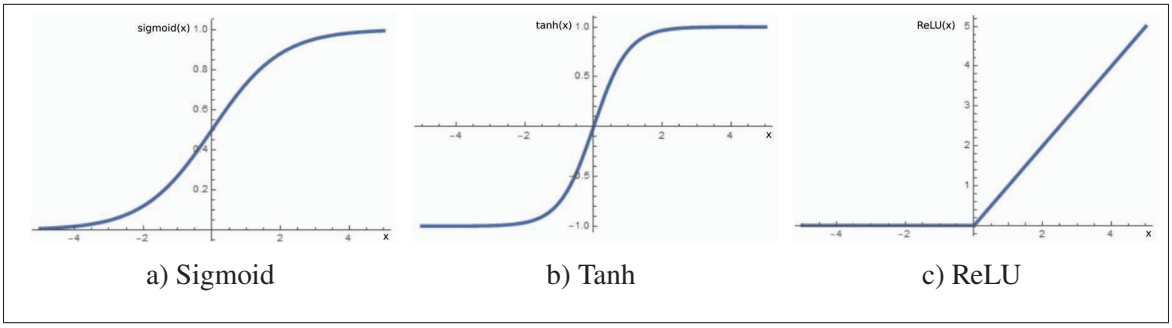
$$\text{ReLU}(x) \;=\; \max(x, 0). \tag{1.3}$$

The curve for ReLU is shown in Figure 1.6 (c). ReLU sets all negative values of x to 0, while keeping positive values unchanged. In other words, it only activates positive inputs and remains inactive for negative inputs. Unlike the sigmoid or tanh function, the ReLU function does not saturate for positive inputs, hence it can alleviate the problem of vanishing gradients in deep neural networks.

As for the softmax function, it is widely used in the output layer of a neural network for multi-class classification problems. This function, which takes a vector of real numbers as input and normalizes them into a probability distribution over multiple classes, is defined as follows

$$[\text{softmax}(\mathbf{x})]_j \;=\; \frac{e^{\mathbf{x}_j}}{\sum_k e^{\mathbf{x}_k}}, \tag{1.4}$$

where j, k are class indexes.



a) Sigmoid b) Tanh c) ReLU

Figure 1.6 Curves of non-linear activation functions.

1.1.2 Loss functions

In deep learning, a loss function, also known as cost function or objective function, is a function quantifying the difference between the predicted output of a neural network and the actual target

output. It plays a crucial role in training the model by providing a feedback signal that guides the model optimization.

The choice of the loss function depends on the task at hand. Popular loss functions in deep learning include the mean squared error (MSE), the cross-entropy, and the Kullback-Leibler (KL) divergence. The MSE loss is defined as

$$\mathrm{MSE}(y, \hat{y}) = \frac{1}{C} \sum_{j=1}^{C} (y_j - \hat{y}_j)^2, \tag{1.5}$$

where $\hat{\mathbf{y}}$ refers to the predicted output of a neural network, $\mathbf{y}$ is the actual target label, and C is the number of output values (e.g., classes for classification). It calculates the mean squared difference between $\mathbf{y}$ and $\hat{\mathbf{y}}$, is sensitive to outliers and penalizes larger errors heavily, which is commonly used for regression tasks.

The cross-entropy loss, which measures the dissimilarity between two probability distributions, is defined as

$$\mathcal{H}(y, \hat{y}) = -\sum_{j=1}^{C} y_j \log \hat{y}_j. \tag{1.6}$$

Here, $\hat{y}_j$ refers to the predicted probability of class j and $\hat{y}_j$ is the true distribution for the same class. Cross-entropy is widely-used in deep learning, particularly for supervised classification and segmentation tasks.

The KL divergence is defined as

$$D_{\mathrm{KL}}\left(y' \,||\, y\right) = \sum_{j=1}^{C} y'_j \log \frac{y'_j}{y_j}, \tag{1.7}$$

where y and y' are two different distributions. Different from cross-entropy that provides a measure of how well the predicted probabilities match the true labels, KL divergence measures the difference in information content or structure between two distributions, and can be used to align or match distributions. We note that KL divergence is not symmetric, i.e.

$D_{\mathrm{KL}}(y' \,||\, y) \neq D_{\mathrm{KL}}(y \,||\, y')$, and is related to cross-entropy as follows

$$D_{\mathrm{KL}}(y' \,||\, y) = \mathcal{H}(y, y') - \mathcal{H}(y), \tag{1.8}$$

where $\mathcal{H}(y) = -\sum_j y_j \log y_j$ is the entropy of y.

1.1.3 Optimization algorithm

Optimization algorithms aim to find the optimal set of parameters that yield the best predictions on the training data by iteratively adjusting the model parameters to minimize the loss function. Various optimization algorithms were developed for training CNNs, such as stochastic gradient descent (SGD) (Harold, Kushner & Yin, 1997), adaptive moment estimation (Adam) (Kingma & Ba, 2014), and rectified Adam (RAdam) (Liu *et al.*, 2019b).

SGD is a fundamental optimization algorithm, which is shown in Algorithm 1.1. It updates the model parameters based on the gradient of the loss function with respect to a mini-batch of training examples. It is computation- and memory-efficiency because the parameter updates are based on a mini-batch rather than the whole dataset. Moreover, mini batches at different iterations are different samples due to the random sampling, leading to slight variations in the computed gradients. This characteristic can help the model escape shallow local minima, allowing it to converge to a better solution. However, it may also introduces oscillations that cause slower convergence.

The Adam maintains adaptive learning rates for each parameter, adjusting them based on the estimates of the first and second moments of the gradients. The details of this optimization algorithm are shown in Algorithm 1.2. Due to the faster convergence and better optimization performance brought by the adaptive learning rate strategy, it is often considered as a default choice for various deep learning architectures.

RAdam algorithm, shown in Algorithm 1.3, is a variant of Adam, which introduces a term to rectify the variance of the adaptive learning rate in the early stage of model training. Compared with Adam, RAdam alleviates performance degradations caused by instability.

Algorithm 1.1 The SGD algorithm

Input: Model $f_\theta(\cdot)$ with initialized parameter θ, training set $\mathcal{D}$ batch size $\mathcal{B}$, and learning rate ϵ
Output: Model parameters θ

while *stopping criterion not met* **do**
  Sample a batch of examples $\mathcal{B}$ from the training set $\mathcal{D} = \{(\mathbf{x}_d, \mathbf{y}_d)\}_{d=1}^{|\mathcal{D}|}$;
  Compute gradient estimate: $\mathbf{g} \leftarrow +\frac{1}{|\mathcal{B}|}\nabla_\theta \sum_{i=1} \mathcal{L}(f_\theta(\mathbf{x}_i), \mathbf{y}_i)$;
  Apply update: $\theta \leftarrow \theta - \epsilon\mathbf{g}$;
end while

Algorithm 1.2 The Adam algorithm

Input: Model $f_\theta(\cdot)$ with initialized parameter θ, stepsize α, exponential decay rates for the moment estimates: $\beta_1, \beta_2 \in [0, 1)$, initial 1^{st} moment vector $m_0 \leftarrow 0$, initial 2^{nd} moment vector $v_0 \leftarrow 0$, and initial time step $t \leftarrow 0$
Output: Model parameters θ_t

while θ_t *not converged* **do**
  $t = t + 1$;
  Sample a batch of examples $\mathcal{B}$ from the training set $\mathcal{D} = \{(\mathbf{x}_d, \mathbf{y}_d)\}_{d=1}^{|\mathcal{D}|}$;
  Compute gradient estimate: $\mathbf{g}_t \leftarrow +\frac{1}{|\mathcal{B}|}\nabla_\theta \sum_{i=1} \mathcal{L}(f_{\theta_t}(\mathbf{x}_i), \mathbf{y}_i)$;
  Update biased first moment estimate: $m_t \leftarrow \beta_1 m_{t-1} + (1 - \beta_1)\mathbf{g}_t$;
  Update biased second moment estimate: $v_t \leftarrow \beta_2 v_{t-1} + (1 - \beta_2)\mathbf{g}_t^2$;
  Correct bias in first moment: $\hat{m}_t \leftarrow \frac{m_t}{(1-\beta_1^t)}$; `/* to the power t */`
  Correct bias in second moment: $\hat{v}_t \leftarrow \frac{v_t}{(1-\beta_2^t)}$;
  Apply update: $\theta_t \leftarrow \theta_{t-1} - \alpha\frac{\hat{m}}{(\sqrt{\hat{v}_t}+\epsilon)}$; `/* small constant ϵ */`
end while

Algorithm 1.3 The RAdam algorithm

Input: Model $f_\theta(\cdot)$ with initialized parameter θ, stepsize α, exponential decay rates for the moment estimates: $\beta_1, \beta_2 \in [0, 1)$, the maximum length of the approximated simple moving average $\rho_\infty \leftarrow \frac{2}{1-\beta_2} - 1$, initial 1^{st} moment vector $m_0 \leftarrow 0$, initial 2^{nd} moment vector $v_0 \leftarrow 0$, and initial time step $t \leftarrow 0$
Output: Model parameters θ_t

while θ_t *not converged* **do**
  $t = t + 1$;
  Sample a batch of examples $\mathcal{B}$ from the training set $\mathcal{D} = \{(\mathbf{x}_d, \mathbf{y}_d)\}_{d=1}^{|\mathcal{D}|}$;
  Compute gradient estimate: $\mathbf{g}_t \leftarrow +\frac{1}{|\mathcal{B}|}\nabla_\theta \sum_{i=1} \mathcal{L}(f_{\theta_t}(\mathbf{x}_i), \mathbf{y}_i)$;
  Update biased first moment estimate: $m_t \leftarrow \beta_1 m_{t-1} + (1 - \beta_1)\mathbf{g}_t$;
  Update biased second moment estimate: $v_t \leftarrow \beta_2 v_{t-1} + (1 - \beta_2)\mathbf{g}_t^2$;
  Correct bias in first moment: $\hat{m}_t \leftarrow \frac{m_t}{(1-\beta_1^t)}$;
  Compute the length of the approximate simple moving average: $\rho_t \leftarrow \rho_\infty - 2t\frac{\beta_2^t}{(1-\beta_2^t)}$;
  if *the variance is tractable, i.e.* $\rho_t > 4$ **then**
    Compute adaptive learning rate: $l_t \leftarrow \sqrt{\frac{1-\beta_2^2}{v_t}}$;
    Compute the variance rectification term: $r_t \leftarrow \sqrt{\frac{(\rho_t-4)(\rho_t-2)\rho_\infty}{(\rho_\infty-4)(\rho_\infty-2)\rho_t}}$;
    Apply update: $\theta_t \leftarrow \theta_{t-1} - \alpha r_t \hat{m}_t l_t$;
  end if
  if *the variance is tractable, i.e.* $\rho_t \leq 4$ **then**
    Apply update: $\theta_t \leftarrow \theta_{t-1} - \alpha \hat{m}_t$
  end if
end while

1.2 Deep learning in medical image segmentation

Different from traditional computer vision techniques for segmentation that relies on handcrafted features (Felzenszwalb & Huttenlocher, 2004; Rother, Kolmogorov & Blake, 2004) and rule-based algorithms (Chan & Vese, 2001; Comaniciu & Meer, 2002), deep learning algorithms can automatically learn hierarchical representations, as well as capture complex patterns and contextual information from images. Neural networks such as CNNs have achieved a remarkable success in various medical image segmentation tasks, such as brain segmentation (Cui *et al.*,

2019; Zhu *et al.*, 2023; Allah, Sarhan & Elshennawy, 2023), cardiac segmentation (Bai *et al.*, 2017; Zotti, Luo, Lalande & Jodoin, 2018; Yu, Wang, Li, Fu & Heng, 2019; Duan *et al.*, 2019; Dong *et al.*, 2020), and abdominal organs segmentation (Wang *et al.*, 2018; Zhou *et al.*, 2019b). In this section, we will introduce some popular network architectures for segmentation, as well as commonly-used loss functions for this task.

1.2.1 Segmentation networks

CNNs are primarily designed for image classification tasks, where the output is a single label indicating the class of the entire image. However, image segmentation requires pixel-level classification, where each pixel in the input image is assigned a label indicating the class to which it belongs. To adapt CNNs to segmentation tasks, the fully convolutional network (FCN) (Long *et al.*, 2015) replaces fully connected layers with standard 1×1 convolutional layers, enabling the network to take an image of arbitrary size and produce dense per-pixel predictions of the same size. An example of a FCN is shown in Figure 1.7. In subsequent years, FCN-based architectures have been further improved, with variations like SegNet (Badrinarayanan, Kendall & Cipolla, 2017), DeepLab (Chen, Papandreou, Kokkinos, Murphy & Yuille, 2014, 2017; Chen, Zhu, Papandreou, Schroff & Adam, 2018), and UNet (Ronneberger, Fischer & Brox, 2015). These architectures are typically composed of an encoder and a decoder, and often incorporate additional techniques such as dilated convolutions, post-processing steps like conditional random fields (CRFs), atrous spatial pyramid pooling (ASPP), and skip-connections to refine the segmentation results.

SegNet (Badrinarayanan *et al.*, 2017), as illustrated in Figure 1.8, has an encoder and a corresponding decoder, where the encoder is a VGG-based architecture (Simonyan & Zisserman, 2014) and the decoder upsamples lower resolution input feature maps using pooling indices to perform non-linear upsampling. The pooling indices are computed correspondingly based on the max-pooling step in the encoder. Then, the upsampled maps are convolved with trainable filters to produce dense feature maps. Compared to the vanilla FCN that uses learnable parameters for upsampling, it reduces the memory requirement for upsampling by storing the indices. However, the upsampling process based on pooling indices may result in limited spatial resolution in the

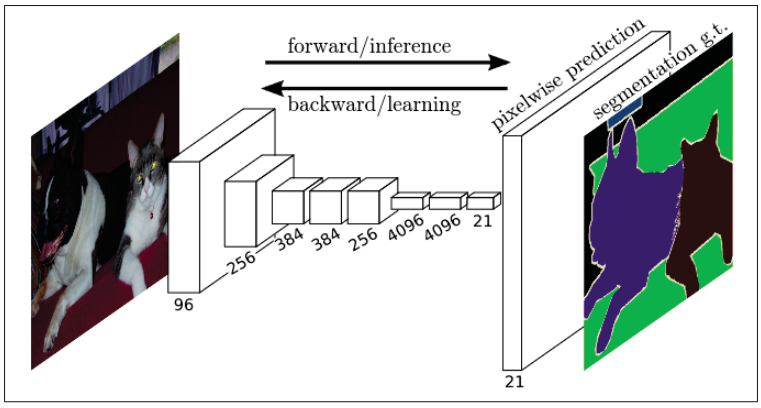


Figure 1.7 A FCN for semantic segmentation. Image is taken from Long *et al.* (2015).

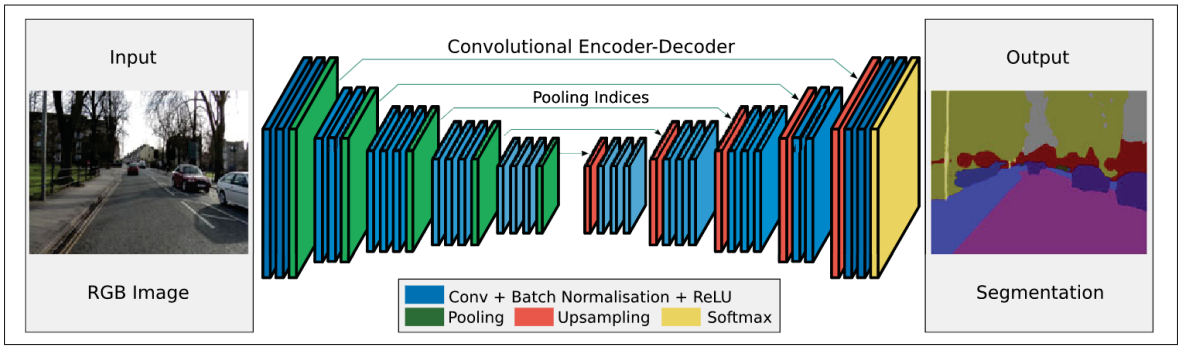


Figure 1.8 SegNet architecture. Image is taken from Badrinarayanan *et al.* (2017).

final segmentation map, losing some fine-grained details that could be important for precise segmentation.

DeepLabV3+ (Chen *et al.*, 2018), shown in Figure 1.9, is the fourth version of DeepLab (Chen *et al.*, 2014). DeepLabV3+, which also has an encoder and a corresponding decoder, uses dilated convolutions in the encoder, allowing the network to capture both local and global contextual information. It also employs ASPP which leverages parallel dilated convolutions at multiple scales to capture multi-scale context effectively, enhancing its ability to segment objects with varying details. In addition, it incorporates skip connections that merge low-level and high-level features from the encoder and decoder stages. This fusion of features enhances the localization accuracy and preserves fine-grained details in segmentation maps. One notable drawback of

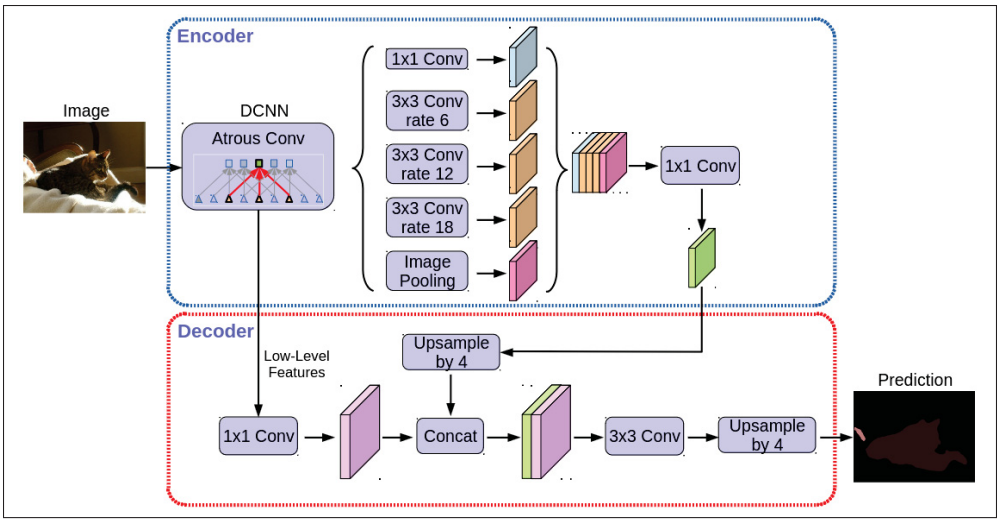


Figure 1.9 DeepLabV3+ architecture. Image is taken from Chen *et al.* (2018).

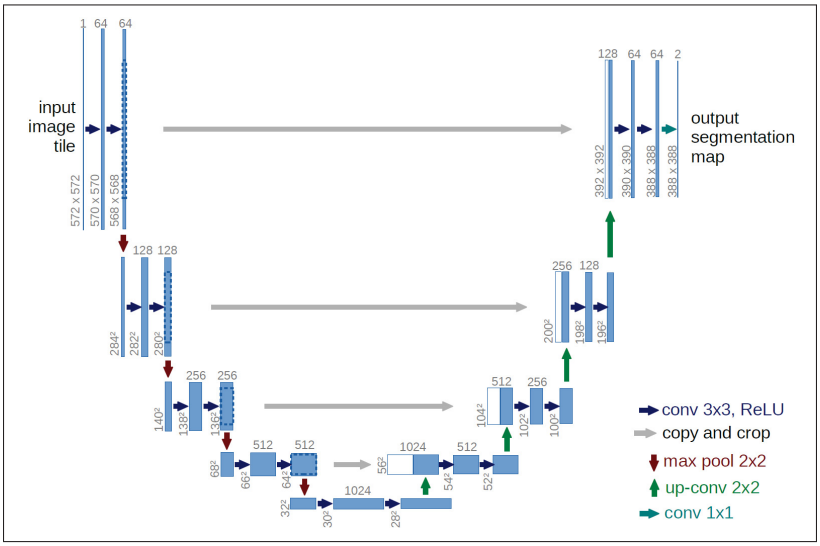


Figure 1.10 UNet architecture. Image is taken from Ronneberger *et al.* (2015).

DeepLabV3+ is its complexity and high computational requirements, which limits its practical applicability.

UNet (Ronneberger *et al.*, 2015) is a highly popular CNN architecture for semantic segmentation tasks, particularly related to medical image analysis. Its architecture is shown in Figure 1.10.

The UNet architecture derives its name from its U-shaped network structure. It consists of an encoder path (contracting path) and a decoder path (expanding path) connected by skip connections. Specifically, the encoder path is responsible for extracting hierarchical features from the input image, and typically consists of multiple convolutional layers followed by a downsampling operation, such as max pooling or strided convolutions. This trick progressively reduces the spatial dimensions while increasing the number of feature channels, capturing high-level representations. On the other hand, the decoder path aims to recover the spatial resolution of the feature maps and generate the segmentation output. It consists of upsampling operations, such as transposed convolutions, to increase the spatial dimensions. The decoder also incorporates skip connections which connect corresponding layers from the encoder path to the decoder path. These connections enable the decoder to leverage low-level and fine-grained information from earlier layers, aiding in precise localization of structures to segment. UNet is known for its ability to handle limited training data effectively and produce high-quality segmentation results, particularly in scenarios with limited annotated samples. Although several variations were proposed, including UNet++ (Zhou, Rahman Siddiquee, Tajbakhsh & Liang, 2018) and Attention UNet (Oktay *et al.*, 2018), the original UNet architecture continues to be a popular and effective choice for semantic segmentation tasks.

Besides UNet, another popular architecture for medical image segmentation is ENet (Adam *et al.*, 2016) which is a highly efficient and lightweight architecture, aiming to achieve a good trade-off between segmentation accuracy and computational efficiency. The architecture of ENet, as presented in Table 1.1, can be divided into several stages. An initial stage that contains a single block is presented in Figure 1.11 a). The encoder includes three stages, where stage 1 consists of 5 bottleneck blocks and stages 2 and 3 have the same structure, with an exception that stage 3 does not have downsample option at the beginning. The decoder consists of stages 4 and 5, along with a final full convolution layer that generates dense predictions. The bottleneck module in ENet is shown in Figure 1.11 b). Both the design of bottleneck blocks with 1×1 and 3×3 convolutions and the lightweight decoder design reduce the number of model parameters and computational cost.

Table 1.1 ENet architecture. Table is taken from Adam *et al.* (2016).

Name	Type	Output size
initial		$16 \times 256 \times 256$
bottleneck1.0	downsampling	$64 \times 128 \times 128$
4× bottleneck1.x		$64 \times 128 \times 128$
bottleneck2.0	downsampling	$128 \times 64 \times 64$
bottleneck2.1		$128 \times 64 \times 64$
bottleneck2.2	dilated 2	$128 \times 64 \times 64$
bottleneck2.3	asymmetric 5	$128 \times 64 \times 64$
bottleneck2.4	dilated 4	$128 \times 64 \times 64$
bottleneck2.5		$128 \times 64 \times 64$
bottleneck2.6	dilated 8	$128 \times 64 \times 64$
bottleneck2.7	asymmetric 5	$128 \times 64 \times 64$
bottleneck2.8	dilated 16	$128 \times 64 \times 64$
Repeat section 2, without bottleneck2.0		
bottleneck4.0	upsampling	$64 \times 128 \times 128$
bottleneck4.1		$64 \times 128 \times 128$
bottleneck4.2		$64 \times 128 \times 128$
bottleneck5.0	upsampling	$16 \times 256 \times 256$
bottleneck5.1		$16 \times 256 \times 256$
fullconv		$C \times 512 \times 512$

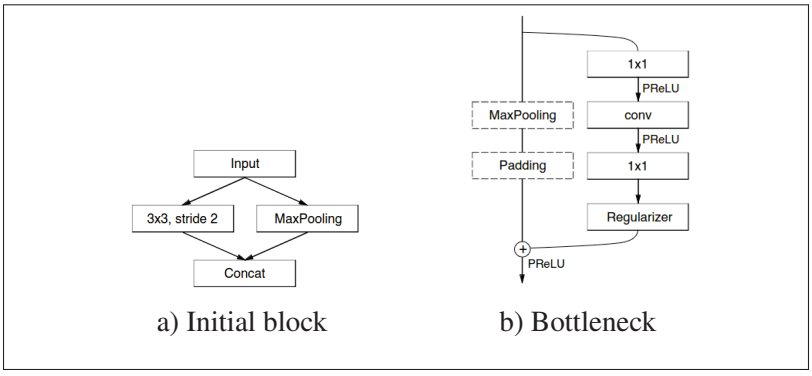


Figure 1.11 Initial block and bottleneck of ENet. Images are taken from Adam *et al.* (2016)

1.2.2 Segmentation loss functions

As previously mentioned, deep learning techniques aim to find the set of model parameters that yield the best predictions on the training data by minimizing the loss function. In this section, we will introduce loss functions that are commonly used for image segmentation. If what follows, we use $i \in \{1, \ldots, N\}$ to denote a pixel index and $j \in \{1, \ldots, C\}$ to denote a class index.

Cross-entropy loss (Yi-de, Qing & Zhi-Bai, 2004) is a popular choice for multi-class segmentation tasks. It measures the dissimilarity between the predicted class probabilities and the one-hot encoded ground truth labels. The cross-entropy loss function defined for multi-class segmentation is

$$\mathcal{L}_{ce}^{seg} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{C} y_{ij} \log \hat{y}_{ij}, \tag{1.9}$$

where $\mathbf{y}$ is the ground truth label encoded as a one-hot vector, $\widehat{\mathbf{y}} = f_{\theta}(\mathbf{x})$ is the predicted class probabilities for image $\mathbf{x}$ produced by the model $f_{\theta}(\cdot)$ with parameters θ. Minimizing the cross-entropy loss encourages the network to assign high probabilities to the correct classes and low probabilities to the incorrect ones.

Focal loss (Lin, Goyal, Girshick, He & Dollár, 2017) is an effective loss when dealing with imbalanced datasets in segmentation tasks. It introduces a modulating factor that down-weights easy examples and focuses more on challenging ones, helping the model prioritize hard-to-segment regions. The focal loss is defined as

$$\mathcal{L}_{focal}^{seg} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{C} (1 - \hat{y}_{ij})^{\gamma} y_{ij} \log \hat{y}_{ij}, \tag{1.10}$$

where $\gamma \geq 0$ is the tunable focusing parameter, and $(1 - \hat{y}_{ij})$ is the modulating factor. If the the pixel is misclassified (i.e., $\hat{y}_{ij}$ is small while $y_{ij} = 1$) then the modulating factor would be near 1 and the loss is almost unaffected. On the other hand, the loss for well-classified examples (i.e., $\hat{y}_{ij}$ is close to 1 while $y_{ij} = 1$) is down-weighted since the modulating factor goes to 0. Another property of focal loss is that γ smoothly adjusts the rate at which easy examples are

down-weighted. When $\gamma = 0$, it is equivalent to cross entropy and, as γ is increased, the effect of the modulating factor is also increased.

Dice loss (Sudre, Li, Vercauteren, Ourselin & Jorge Cardoso, 2017) is another popular loss function used in image segmentation tasks. It measures the overlap or similarity between the predicted segmentation mask and the ground truth mask. This loss, which is particularly effective for tasks where class imbalance is a challenge and is defined as follows:

$$\mathcal{L}_{Dice}^{seg} = 1 - \frac{1}{C}\sum_{j=1}^{C}\frac{2\sum_i y_{ij}\hat{y}_{ij}}{\left(\sum_i y_{ij} + \sum_i \hat{y}_{ij} + \epsilon\right)}. \tag{1.11}$$

The numerator in the right term estimates the number of positive pixels between the predicted $\widehat{\mathbf{y}}$ and ground truth masks $\mathbf{y}$, while the denominator computes the sum of positive pixels in both masks, and ϵ is a small constant to avoid division by zero. Minimizing the Dice loss encourages the model to correctly identify the boundaries and spatial extent of target regions in the predicted segmentation.

1.3 Self-paced learning

Self-paced learning (SPL) originates from a human learning technique that allows individuals to define their learning speed according to their unique learning patterns. For instance, when a person first begins learning about mathematics, this person initially focuses on counting, then progressed to addition and subtraction. It is not until a certain age that more advanced concepts like matrix multiplication are encountered. Similarly, in the context of SPL for deep learning, the approach involves starting with relatively easy examples, and once learned, it continues with harder ones benefiting from the already learned knowledge. The general formulation of SPL in deep learning can be described as follows (Ma, Meng, Dong & Yang, 2020):

$$\min_{\theta,\mathbf{v}\in\mathbb{R}^N} \sum_{i=1}^{N} v_i \,\mathcal{L}\big(\mathbf{y}_i, f_\theta(\mathbf{x}_i)\big) \;+\; \mathcal{R}_\gamma(\mathbf{v}). \tag{1.12}$$

This formulation has two main differences compared to supervised training. First, the loss for each example $(\mathbf{x}_i, \mathbf{y}_i)$ is weighted by a variable $v_i \in [0, 1]$ which controls its importance in the learning objective. These learning weights are optimized jointly with the parameters θ of the network. Second, a term $\mathcal{R}_\gamma(\mathbf{v})$ called self-paced regularizer is added to the objective. The role of this term is to control to learning weights $\mathbf{v}$ based on a learning pace parameter γ, so that the model can learn from easy examples before learning harder ones. Toward this goal, the self-paced regularizer must satisfy three conditions (Lu, Meng, Zhao, Shan & Hauptmann, 2015): 1) $\mathcal{R}_\gamma(\mathbf{v})$ is convex with respect to $\mathbf{v}$, which ensures the soundness for optimization; 2) v_i is monotonically *decreasing* with respect to the loss for $(\mathbf{x}_i, \mathbf{y}_i)$, i.e., harder examples should have a smaller weight; 3) v_i is monotonically *increasing* with respect to γ, i.e., increasing the learning pace also increases the weights of examples until they reach a value of 1. Using an alternating optimization strategy that iteratively updates $\mathbf{v}$ and θ while gradually increasing the learning pace γ, harder samples can be automatically included into training.

1.4 Learning from non-differentiable losses

Some applications of deep learning require to handle non-differentiable objectives during training or inference. This situation often occurs in reinforcement learning algorithms where discrete actions are samples from a distribution. Non-differentiable losses can also arise when the desired property for a prediction cannot be measured directly with a numerical function and instead requires running an algorithm (e.g., convexity of a region).

The Reinforce algorithm (Williams, 1992), also known as the score function estimator (Fu, 2006), has been used extensively in reinforcement learning to deal with this problem. This algorithm works as follows. Given a random variable $x \sim p_\theta(x)$ where $p_\theta(x)$ is a parametric distribution, and a function J for which we wish to compute the gradient of its expected value,

$$\nabla_\theta \, \mathbb{E}_{x \sim p_\theta(x)}[J(x)] \tag{1.13}$$

The algorithm uses a simple log-derivative trick to simplify the differentiation process:

$$\nabla_\theta p_\theta(x) = p_\theta(x) \nabla_\theta \log p_\theta(x). \tag{1.14}$$

Using this trick, Eq.(1.13) can be reformulated as

$$\begin{aligned} \nabla_\theta \, \mathbb{E}_{x \sim p_\theta(x)}[J(x)] &= \nabla_\theta \int J(x) p_\theta(x) dx \\ &= \int J(x) \nabla_\theta p_\theta(x) dx \\ &= \int J(x) p_\theta(x) \nabla_\theta \log p_\theta(x) dx \\ &= \mathbb{E}_{x \sim p_\theta(x)}[J(x) \nabla_\theta \log p_\theta(x)] \end{aligned} \tag{1.15}$$

Assuming it is possible to cheaply sample from the distribution $p_\theta(x)$, Monte Carlo sampling is used to approximate the expectation, as follows:

$$\begin{aligned} \nabla_\theta \, \mathbb{E}_{x \sim p_\theta(x)}[J(x)] &= \mathbb{E}_{x \sim p_\theta(x)}[J(x) \nabla_\theta \log p_\theta(x)] \\ &\approx \frac{1}{N} \sum_{i=1}^{N} J(x_i) \nabla_\theta \log p_\theta(x_i) \end{aligned} \tag{1.16}$$

As function J is not derived directly, the Reinforce algorithm places no restriction on the nature of this function, allowing it to be non-differentiable.

1.5 Medical image segmentation scenarios

In this section, we introduce two important scenarios of medical image segmentation, related to semi-supervised learning and unsupervised domain adaptation.

1.5.1 Semi-supervised image segmentation

We start by defining the task of semi-supervised segmentation. Let $\mathcal{S} = \{(\mathbf{x}_s, \mathbf{y}_s)\}_{s=1}^{|\mathcal{S}|}$ be a small amount of labeled data and $\mathcal{U} = \{\mathbf{x}_u\}_{u=1}^{|\mathcal{U}|}$ be a large amount of unlabeled data, where

$\mathbf{x}_s, \mathbf{x}_u \in \mathbb{R}^{|\Omega|}$ are images, $\mathbf{y}_s \in \{0,1\}^{|\Omega|\times|C|}$ is the corresponding labels of labeled image $\mathbf{x}_s$. Here, Ω denotes the set of image pixels and C the set of segmentation classes. The goal of semi-supervised segmentation is to leverage both labeled data $\mathcal{S}$ and unlabeled data $\mathcal{U}$ for improving the performance of a segmentation model $f_\theta(\cdot)$ and reduce the burden of acquiring annotations.

Various semi-supervised methods have been proposed to boost the accuracy and generalization of medical image segmentation models. These methods can be roughly classified into five categories: self-training based methods (Lee *et al.*, 2013; Zheng *et al.*, 2020; Xie, Luong, Hovy & Le, 2020), entropy-based methods (Grandvalet & Bengio, 2004; Hang *et al.*, 2020), consistency regularization methods (Laine & Aila, 2016; Bortsova, Dubost, Hogeweg, Katramados & De Bruijne, 2019; Antti & Valpola, 2017; Miyato *et al.*, 2019), adversarial learning methods (Luc, Couprie, Chintala & Verbeek, 2016; Souly, Spampinato & Shah, 2017; Zhang *et al.*, 2017; Hung, Tsai, Liou, Lin & Yang, 2018; Zhang, Li, Zhang & Ma, 2020), and co-training based methods (Zhou *et al.*, 2019b; Peng *et al.*, 2020a; Xia *et al.*, 2020a; Zheng *et al.*, 2022).

Self-training is one of the earliest approaches for semi-supervised learning (Scudder, 1965; Fralick, 1967). In this approach, an initial segmentation model is constructed by using labeled data $\mathcal{S}$, and then this model is used to predict labels (pseudo labels) for unlabeled data from $\mathcal{U}$. The learning process of the initial model is continued with the newly labeled examples that have been determined to be correctly labeled based on a selection criterion. Pseudo-Label (Lee *et al.*, 2013) is a classical self-training method where the segmentation model is trained in a supervised manner using $\mathcal{S}$ and $\mathcal{U}$ simultaneously. For unlabeled data, the classes which have the maximum predicted probability are selected as pseudo labels. For this kind of method, the quality of the pseudo labels is crucial as bad labels may lead to an even worse segmentation. To further guarantee the quality of generated pseudo labels, Zheng *et al.* (Zheng *et al.*, 2020) proposed a self-training framework which can automatically estimate and refine the pseudo labels, selecting only those well labeled samples to expand training set for retraining the segmentation model. Besides the segmentation network, the framework consists of two modules: 1) a quality estimation network that filters the pseudo labels based on a shape confidence score generated

by an auxiliary variational auto-encode (VAE) and a semantic matching score produced by a VGG network; 2) a refinement network which is used to improve the filtered pseudo labels by adversarial learning. Another recently proposed self-training method (Xie *et al.*, 2020) address inaccurate predictions of unlabeled data in the early training stage. Precisely, a teacher model is first trained on labeled data and then used to generate pseudo labels for unlabeled images. Afterwards, a larger student model is trained on the combination of labeled and pseudo labeled images with injected noise. This process is repeated iteratively, using the student at the previous iteration as the teacher in the current one. In the paper, it is shown that three iterations are sufficient to converge. While these methods achieve a competitive performance, they are prone to collapse when the limited amount of labeled data leads to incorrect pseudo labels.

Entropy-based methods encourage the model to produce more confident predictions by adding a penalty based on the entropy of the predicted probability distribution. This method promotes the model to assign higher probabilities to correct classes and lower probabilities to incorrect ones. Grandvalet and Bengio (Grandvalet & Bengio, 2004) incorporated unlabeled data in the standard supervised learning by introducing an entropy minimization regularization loss term. Hang *et al.* (Hang *et al.*, 2020) proposed a structure-aware entropy regularized semi-supervised method, which applies entropy minimization on student output and encourages the consistency of inter-voxel similarities within the same local region of predictions from teacher and student networks, to improve the generalization of neural networks for left atrium segmentation. Entropy has also been used for selecting confident predictions or as a regularization term in other semi-supervised approaches such self-training (Xie *et al.*, 2020; Reed *et al.*, 2014), adversarial learning (Mugnai, Pernici, Turchini & Del Bimbo, 2022; Vu, Jain, Bucher, Cord & Pérez, 2019), and co-training (Wang *et al.*, 2021a; Shen *et al.*, 2023).

Consistency regularization methods are a popular and important line of work in semi-supervised learning. This kind of method is based on the transform-invariant principle that a model's predictions should be invariant to certain transformations applied to the input data, encouraging the model to produce consistent predictions on perturbed versions of the same input. Laine and Aila (Laine & Aila, 2016) proposed a Π-model to implement self-ensembling, which enforces

consistent predictions for two perturbed versions of the same unlabeled image via different dropout and augmentations such as Gaussian noise. Based on the idea of self-ensembling from (Laine & Aila, 2016), Bortsova *et al.* (Bortsova *et al.*, 2019) proposed to enforce the segmentation network to learn consistency under transformation, successfully boosting the accuracy of chest X-ray images segmentation. Tarvainen and Valpola (Antti & Valpola, 2017) developed a well-known semi-supervised learning method called Mean Teacher. In addition to ensuring consistent predictions between the student and teacher models when subjected to various types of noise, this method updates the teacher model's weights using an exponential moving average of the student weights. This update mechanism, which incurs few additional computations, allows the teacher model to adapt gradually and incorporate the knowledge learned by the student over time, leading to improved performance. Instead of using random perturbations such as dropout, Gaussian noise, and transformations, enforcing consistency with adversarial examples generated through predictable and interpretable perturbations is another promising way to improve the robustness of models (Miyato *et al.*, 2019). In this approach, the perturbations are constrained to a given subspace and estimated by maximizing the model's disagreement on perturbed samples. The adversarial examples are produced by adding the estimated perturbations on the original unlabeled data. Then, the consistency between the prediction of the original image and the adversarial example is minimized, encouraging the model to become more robust.

Adversarial learning is another popular approach for semi-supervised segmentation, which benefits from the work on generative adversarial networks (GANs) (Goodfellow *et al.*, 2020). This kind of method relies on a discriminator that predicts whether the output of the segmentation network for labeled and unlabeled examples belongs to the same distribution. Based on this principle, Luc *et al.* (Luc *et al.*, 2016) trained a segmentation network alone with an adversarial network to detect and correct higher-order inconsistencies between the ground truth and the predicted segmentation maps. Zhang *et al.* (Zhang *et al.*, 2017) proposed a three-stage adversarial training framework in which a segmentation network is first trained with labeled data, after which an evaluation network is constructed. This evaluation network outputs scores

for predictions of the segmentation network for both labeled images and unlabeled images. Finally, the segmentation network is further trained along with the evaluation network in an adversarial manner using unlabeled data. Instead of predicting whether a sample belongs to the data distribution, some adversarial approaches for semi-supervised segmentation employ a fully-convolutional discriminator to make a prediction at each pixel. Following this idea, Souly *et al* (Souly *et al.*, 2017) proposed to include generated data in the training to force labeled and unlabeled samples to be close in the feature space. In the proposed approach, the discriminator is a pixel-wise multi-class classifier which is used to output a $|C|+1$-class probability map, with $|C|$ semantic classes and additional class for fake images. Hung *et al.* (Hung *et al.*, 2018) used a pixel-wise discriminator trained with labeled data to compute a confidence map for self-training. For unlabeled images, the high-confidence predictions are used as pseudo-labels to train the segmentation network. To further improve the generalization and stability of the network, Zhang *et al.* (Zhang *et al.*, 2020) introduced self-attention modules to the segmentation network, and applied a spectral normalization technique to the discriminator.

Co-training based methods assume that different views of data provide complementary information (Blum & Mitchell, 1998). The objective of such methods is to train multiple models on various data views, enabling them to exchange and learn from each other's predictions. Zhou *et al.* (Zhou *et al.*, 2019b) trained three models for multi-planar images of 3D CT volumes (sagittal, axial, and coronal planes), and used the aggregated outputs from the models to guide the learning. Similarly, Xia *et al.* (Xia *et al.*, 2020a) trained multiple models using pseudo labels on multi-view data generated through various transforms. The pseudo labels for the output of current model are estimated by aggregating predictions from the other models based on an epistemic uncertainty. Zheng *et al.* (Zheng *et al.*, 2022) proposed an uncertainty-aware co-training method to make models learn high-confidence regions. Specifically, they used Monte Carlo Sampling to estimate an uncertainty map that serves as a weight for consistency losses between co-trained models, and forcing the models to focus on the valuable region. Peng *et al.* (Peng *et al.*, 2020a) proposed a co-training framework for semi-supervised segmentation, where two models are trained with labeled data, and unlabeled images are used to exchange

information with each other via an agreement loss and a diversity loss. The diversity loss, which serves to increase the diversity across models, is calculated from adversarial samples.

Constrained based methods Despite their ability to boost performance for segmentation when few labeled images are available, the aforementioned methods may be unable to learn the distribution of valid shapes, impeding their application in real-life clinical scenarios. To generate anatomically-plausible predictions, other approaches have proposed to incorporate segmentation constraints in a semi-supervised setting. For example, Zotti *et al.* (Zotti *et al.*, 2018) used labeled data to learn a 3D statistical map measuring the probability that a given voxel belongs to a specific class. Zhou *et al.* (Zhou *et al.*, 2019a) proposed a prior-aware neural network that constrains the relative sizes of target regions in the output, ensuring their proximity to a prior derived from the labeled dataset. Another approach guides both the location and shape of segmented regions using a set of annotated anatomical atlases (Zheng *et al.*, 2019; Duan *et al.*, 2019; Dong *et al.*, 2020). Other works (Oktay *et al.*, 2017; Gao *et al.*, 2021; Painchaud *et al.*, 2020) employed the reconstruction error of an auto-encoder on the predicted segmentation as a prior to guide the segmentation. Ganaye *et al.* Ganaye, Sdika & Benoit-Cattin (2018) considered the inherent nature of anatomical structures within brain regions and introduced a connectivity constraint to improve the robustness of the segmentation. The connectivity constraint is implemented by using an adjacency graph estimated from the ground-truth segmentation masks.

Most of these methods require a good amount of annotated images to compute the shape prior, which may not be possible in semi-supervised settings. In view of this limitation, constraint-based methods that do not need labeled images have also been developed (Pathak, Krahenbuhl & Darrell, 2015; Kervadec *et al.*, 2019; Jia, Huang, Eric, Chang & Xu, 2017). Pathak *et al.* (Pathak *et al.*, 2015) incorporated inequality constraints in a segmentation network, enforcing these constraints on a latent distribution instead of directly on the network output. This enables decoupling the SGD-based optimization of network parameters and the constrained optimization problem on the latent distribution. In contrast, Kervadec *et al.* (Kervadec *et al.*, 2019) modeled inequality constraints on the size of target regions directly in the loss function. This provides a simpler optimization compared to the Lagrangian-based constrained CNNs

in (Pathak *et al.*, 2015), which yielded better results. Jia *et al.* (Jia *et al.*, 2017) used a simple L2 penalty to impose equality constraints on the size of the target regions in the task of histopathology image segmentation. While these methods provide a more plausible segmentation, they are limited to simple constraints that cannot fully characterize the complex shapes found in medical imaging applications. How to incorporate complex non-differentiable anatomical constraints on a segmentation network to generate anatomical-plausible predictions is still an open question.

1.5.2 Domain adaptation for medical image segmentation

In contrast to the semi-supervised scenario where labeled and unlabeled data originate from the same distribution, domain adaptation tackles a more challenging setting where the labeled and unlabeled data stem from different distributions. The UDA scenario for the task of segmentation can be defined as follows. Let $\mathcal{S} = \{(\mathbf{x}_s, \mathbf{y}_s)\}_{s=1}^{|\mathcal{S}|}$ be labeled data from a source domain and $\mathcal{T} = \{\mathbf{x}_t\}_{t=1}^{|\mathcal{T}|}$ unlabeled data from a target domain, where the source and target domains have different distributions. The goal of UDA is to train a segmentation model $f_\theta(\cdot)$ using $\mathcal{S}$ and $\mathcal{T}$ so that it gives accurate predictions for new images of the target domain.

UDA method can be broadly divided into two categories based on their primary focus, domain shift reduction methods (Kumagai & Iwata, 2019; Wang, Li, Ding & Wang, 2020; Ganin *et al.*, 2016; Chen, Dou, Chen, Qin & Heng, 2020) and domain-invariant representation methods (Yang *et al.*, 2019a; Dai *et al.*, 2021a; Bateson *et al.*, 2020).

Domain shift reduction methods aim to reduce the domain shift between the source and target domains by aligning the features or visual styles of domains. The goal is to make the source and target domains more similar, allowing the segmentation model trained on the source domain to generalize well on the target domain. In (Kumagai & Iwata, 2019; Wang *et al.*, 2020), intermediate features distributions are aligned across domains by minimizing their maximum mean discrepancy (MMD). This direct method is however prone to align irrelevant or noisy features that do not contribute to learning the task. This can lead to a reduction in the discriminative capacity of model, resulting in decreased performance on the target domain. Ganin

et al. (Ganin *et al.*, 2016) used an adversarial learning technique to train a neural network, aiming to map the intermediate feature maps from source and target examples towards a representation space where they become indistinguishable. Chen *et al.* (Chen *et al.*, 2020) achieved synergistic alignment of domains from both image and feature space with a style transfer strategy based on CycleGANs. However, such adversarial method requires solving a hard minimax optimization problem which makes the training unstable and can suffer from the problem of mode collapse.

Domain-invariant representation learning aims to learn representations that are applicable to both source and target domains. The goal is to extract features that capture the underlying semantics of the data while being less affected by domain-specific variations. Instead of using pixel-wise style transfer models like CycleGANs, Yang *et al.* (Yang *et al.*, 2019a) preserved semantic feature-level information by finding a shared content space. They first encode images from each domain into two different feature spaces, a domain-invariant content space and a domain-specific style space. Then, the representation in the domain-invariant content space is extracted to perform the segmentation task. This method usually depends on a reconstruction task with an auxiliary decoder. To avoid any additional generative model or decoder, Dai *et al.* (Dai *et al.*, 2021a) proposed a dynamic task-oriented disentangling network to learn disentangled representations that include task-relevant representation associated with the critic semantic information and task-irrelevant representation encoded the domain-specific information. These representations are disentangled by regularizing a group of task-specific loss functions. Bateson *et al.* (Bateson *et al.*, 2020) proposed a novel formulation for adaptation by introducing domain-invariant priors that derived from anatomical information. They integrated a class-ratio prior into the segmentation network by minimizing KL divergence between the class marginal distribution for target examples and a reference empirical distribution estimated on source examples. While this simple approach improves performance, it only considers shape size as prior. However, other priors that capture spatial relations among classes can also be useful for domain adaptation.

1.6 Conclusion

In this chapter, we presented a comprehensive overview of key concepts and background knowledge in deep learning, and highlighted their application in the field of medical image segmentation. Furthermore, we introduced two challenging scenarios in medical image segmentation, semi-supervised segmentation and unsupervised domain adaptation, and provided a detailed description of the main approaches proposed for addressing these scenarios.

Despite the significant progress made in semi-supervised segmentation and unsupervised domain adaptation for segmentation, there are still limitations that need to be addressed. In the case of semi-supervised segmentation, where a small number of labeled images and a large amount of unlabeled images are used for training the network, there is a tendency for inaccurate predictions to occur during the initial training stage, particularly for challenging unlabeled samples. As the training progresses, this initial inaccuracy in predictions can be reinforced, especially for methods based on self-training, leading to a deterioration in performance over time. On the other hand, although semi-supervised segmentation methods can achieve high performance in terms of objective metrics like the Dice score, they may generate predictions that are deemed anatomically implausible by clinicians. For instance, these methods might produce non-connected predictions that should be considered a single region from an anatomical perspective or non-convex regions that are expected to be convex. To address this issue and prevent anatomically impossible segmentations, it is crucial to incorporate anatomical priors into the network. However, incorporating such priors is often challenging since they are typically non-differentiable and therefore cannot be easily integrated into a loss function directly. In the context of domain adaptation segmentation, previous approaches usually relied on auxiliary reconstruction tasks or style transfer tasks, leading to complex pipelines. A simpler and highly effective domain adaptation method proposed by Bateson *et al.* (Bateson *et al.*, 2020) directly aligns domain-invariant information using an objective function, achieving competitive performance without the need for additional reconstruction decoders or adversarial learning. However, the current use of a domain-invariant prior based solely on shape size presents limitations. There is a pressing need to explore priors that encode richer characteristics of

the structures to segment. The following chapters of the thesis present novel semi-supervised learning and unsupervised domain adaptation methods for medical image segmentation that address these limitations.

CHAPTER 2

SELF-PACED AND SELF-CONSISTENT CO-TRAINING FOR SEMI-SUPERVISED IMAGE SEGMENTATION

Ping Wang[1] , Jizong Peng[1] , Marco Pedersoli[1] , Yuanfeng Zhou[2] , Caiming Zhang[2] , Christian Desrosiers[1]

[1] Department of Software and IT Engineering, École de Technologie Supérieure, 1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3

[2] School of Software, Shandong University, 1500 Middle of Shunhua Road, Jinan, Shan Dong, China 250101

Article published in Journal « Medical Image Analysis » in June 2021.

2.1 Introduction

Semi-supervised learning, where the goal is to learn a given task with few labeled examples and many unlabeled ones, has generated growing interest in research. This learning paradigm is of key importance for medical imaging (Cheplygina, de Bruijne & Pluim, 2019) since obtaining annotated data in applications of this field is often costly and typically requires highly-trained experts. In the recent years, a broad range of deep learning approaches have been proposed for semi-supervised learning. A method which has gained popularity is co-training. Initially proposed by Blum and Mitchell for classification (Blum & Mitchell, 1998), this method exploits the idea that training examples can often be described by two complementary sets of features called views. It was shown that models trained collaboratively on conditionally-independent views improve semi-supervised performance with PAC-style bounds on the generalization error (Dasgupta, Littman & McAllester, 2002).

The standard dual-view co-training approach consists in learning a separate classifier for each view using labeled data. Information is then exchanged between classifiers based on their high-confidence predictions for unlabeled data. The generalization of this approach to more than two views is called multi-view learning (Xu, Tao & Xu, 2013). By enforcing agreement between classifiers, multi-view learning constrains the parameter search space and helps find

models which can generalize to unseen data. Despite its success in various classification tasks, its application to image analysis problems has so far been limited, mostly due to the requirement of having independent features in each view. Although these complementary views are available in specific scenarios such as multiplanar images (Zhou *et al.*, 2019b; Xia *et al.*, 2020a), there is no effective way to construct them for arbitrary images. Qiao *et al.* (Qiao, Shen, Zhang, Wang & Yuille, 2018) proposed a deep co-training method for semi-supervised image recognition, where adversarial examples built from training images are used to enforce diversity among different classifiers. This idea was later extended to medical image segmentation by (Peng *et al.*, 2020a). More recently, Xia *et al.* (Xia *et al.*, 2020a) introduced an uncertainty-aware multi-view co-training framework in which prediction uncertainty estimated using a Bayesian approach is employed to merge the output for different views.

Despite improving performance when few labeled images are available, existing co-training approaches for semi-supervised segmentation suffer from two important limitations: 1) they do not employ a self-paced learning strategy and therefore are susceptible to incorrect predictions in initial stages of training; 2) they do not exploit self-consistency within the model. In this paper, we address these limitations by introducing a self-paced and self-consistent co-training method. The main contributions of our work are as follows:

- We propose, to our knowledge, the first self-paced co-training method for semi-supervised segmentation. We show that the proposed self-paced learning strategy corresponds to minimizing a generalized Jensen-Shannon Divergence (JSD), where the confidence of individual models for each pixel in unlabeled images is considered. This enables the learning process to focus on transferring high-confidence predictions across models, thereby boosting overall accuracy;
- Our method is also the first to incorporate self-consistency based on temporal ensembling directly in a co-training framework. We demonstrate empirically that self-paced learning and self-consistency regularization play a complementary role in semi-supervised segmentation, and that their combination leads to state-of-the-art performance;

- We also present an uncertainty-regularized version of our self-paced JSD loss, which further leverages entropy regularization to enforce joint confidence across the different models. By doing so, confident predictions for unlabeled images are learned better;
- We perform one of the most comprehensive analyses of co-training for semi-supervised medical image segmentation, evaluating the proposed method on three different segmentation tasks and against five recently-proposed approaches for this problem. Results demonstrate the advantages of the proposed method compared to existing approaches; The code is available at https://github.com/WangPing521/self-paced-and--self-consistent-semi-supervised-segmentation.

The rest of this paper is structured as follows. In Section 2.2, we give an overview of relevant literature on semi-supervised segmentation and related work on entropy regularization and self-paced learning. Section 2.3 then describes our self-paced and self-consistent co-training method for semi-supervised segmentation. Afterwards, Section 4.4 and 4.5 present the experimental setup and obtained results. We conclude the paper with a summary of our work's main contributions and its potential extensions in Section 2.6.

2.2 Related work

2.2.1 Semi-supervised segmentation

The bulk of semi-supervised methods for segmentation can be roughly grouped into four categories: self-training methods (Zou, Yu, Kumar & Wang, 2018; Bai *et al.*, 2017), regularization methods (Chaitanya *et al.*, 2019; Zhao, Balakrishnan, Durand, Guttag & Dalca, 2019; Bortsova *et al.*, 2019; Cui *et al.*, 2019; Perone, Ballester, Barros & Julien, 2019; Yu *et al.*, 2019; Dou, Liu, Heng & Glocker, 2020), adversarial learning (Souly *et al.*, 2017; Zhang *et al.*, 2017; Hung *et al.*, 2018; Mondal, Dolz & Desrosiers, 2018), and co-training methods (Peng *et al.*, 2020a; Xia *et al.*, 2020a; Zhou *et al.*, 2019b).

In basic self-training approaches, a model generates pseudo-labels for unlabeled data and is then retained with the updated set of labeled examples. The Pseudo-label algorithm proposed by Lee (Lee *et al.*, 2013) fine-tunes the model with the new pseudo-labeled data instead of retraining it at each pseudo-labeling step. Since pseudo-labels predicted in earlier training stages are generally less reliable, their importance in the loss function is gradually increased over training. When annotated data is very scarce, incorrect predictions may however be reinforced by this approach, leading to a worse performance.

A wide range of regularization-based methods have also been proposed for semi-supervised classification and segmentation. The Γ model (Rasmus *et al.*, 2015) evaluates unlabeled data samples with and without noise, and then applies a consistency loss between the two predictions. However, if the model predicts incorrect targets, enforcing consistency on wrong predictions may impede learning. To mitigate this problem, Miyato *et al.* (Miyato *et al.*, 2019) proposed a virtual adversarial training (VAT) regularization method where a divergence-based local smoothness loss is employed to make the model robust to adversarial perturbations of the input. Laine *et al.* (Laine & Aila, 2016) presented a knowledge distillation method called temporal ensembling which encourages consistent network outputs for different evaluations and dropout conditions of the same input. This is achieved by aggregating the predictions of multiple previous network evaluations into an ensemble (the Teacher), and minimizing the L_2 distance between predictions of the ensemble and the current model (the Student). A drawback of this approach is that the learned information is incorporated into the training process at a slow pace since each target is updated only once per epoch. To overcome this limitation, Antti *et al.* (Antti & Valpola, 2017) developed the Mean Teacher method, which averages and compares model weights instead of predictions. In addition, since weight averages involve all layers, not only the last one, the target model can learn better intermediate representations.

Recently, several semi-supervised segmentation methods were proposed based on knowledge distillation approaches like Mean Teacher (Cui *et al.*, 2019; Perone *et al.*, 2019; Yu *et al.*, 2019; Dou *et al.*, 2020). Yu *et al.* (Yu *et al.*, 2019) presented an Uncertainty-Aware Mean Teacher (UA-MT) framework where Monte-Carlo dropout is employed to estimate the pixel-

wise prediction uncertainty of the teacher, and uncertainty values are used as weights in the consistency loss between the teacher and student outputs (i.e., output consistency for the teacher's confident predictions is given more importance). While our method also exploits confidence to transfer knowledge across models, there are important differences compared to this existing approach. Whereas UA-MT estimates uncertainty in an extrinsic fashion (i.e., with Monte Carlo dropout sampling) and for a single teacher network, the proposed model considers the prediction confidence of multiple networks directly in the objective function. Specifically, our model includes self-paced learning weights which control the reliability of a pseudo-label for a given pixel as separate variables in the learning process, and solves a global optimization problem including both these weights and network parameters. In contrast, UA-MT relies on a manually-selected threshold on the dropout uncertainty to select confident predictions. Another important difference with the work of (Yu *et al.*, 2019) is that our method also incorporates an uncertainty regularization loss that encourages the trained networks to both agree with each other and be confident in their prediction.

Whereas these approaches are agnostic to the segmentation task, Chaitanya *et al.* (Chaitanya *et al.*, 2019) proposed a data augmentation method where a generative model is trained with task-specific data to generate realistic images and segmentation masks. Similar techniques based on transformation consistency are presented in (Zhao *et al.*, 2019; Bortsova *et al.*, 2019). In our experiments, we show that the proposed method outperforms state-of-the-art regularization-based approaches for semi-supervised segmentation.

Adversarial learning methods for semi-supervised segmentation typically use a classification network (the discriminator) during training to predict if the output of the segmentation network (the generator) is from the same or different distribution compared to labeled examples (Zhang *et al.*, 2017; Luc *et al.*, 2016; Souly *et al.*, 2017; Mondal *et al.*, 2018). By trying to fool the discriminator, the generator is encouraged to output a similar predictive distribution for both images with and without annotations. A potential issue with this approach is that the adversarial network can have a reverse effect, where the output for annotated images becomes increasingly similar to the wrong predictions for unlabeled images. A related strategy proposed

by (Hung *et al.*, 2018) uses the prediction of a fully-convolutional discriminator at each pixel as a confidence map for the segmentation. For unlabeled images, predictions in high-confidence regions are then considered as pseudo-labels to update the segmentation network. Despite their success, adversarial learning approaches are often avoided due to the complexity and instability of their training.

Co-training methods have also shown promising results for semi-supervised segmentation. Peng *et al.* (Peng *et al.*, 2020a) introduced a deep co-training method which combines a consistency loss based on JSD and a model diversity loss using adversarial training. Zhou *et al.* (Zhou *et al.*, 2019b) used different planes of a 3D MRI scan as separate co-training views and use the aggregated prediction on unlabeled images to guide the learning. In their paper, Xia *et al.* (Xia *et al.*, 2020a) extended this last framework by considering Bayesian-estimated uncertainty when merging the predictions of different views. While this approach considers prediction uncertainty, it does so without taking into account learning pace. In comparison, our proposed method provides a principled technique for self-paced co-training based on a generalized JSD, where high-confidence predictions are leveraged in a dynamic fashion to co-regularize the segmentation networks. To our knowledge, our work is also the first to propose self-consistency regularization within co-training.

2.2.2 Entropy regularization

Entropy minimization was first proposed in (Grandvalet & Bengio, 2004) to improve learning in semi-supervised classification. The basic idea of this approach is encouraging a model to have confident predictions for unlabeled examples by minimizing their entropy. This forces the decision boundary to go through a low-density region of the data, thereby helping the classifier generalize to unseen examples. Although this approach has shown great potential for unsupervised domain adaptation (Vu *et al.*, 2019), its application to semi-supervised segmentation remains limited. In this work, we extend the concept of semi-supervised entropy regularization, which has until now been used in a single-model scenarios, to the more general multi-view co-training setting.

2.2.3 Self-paced learning

Self-paced learning (SPL) (Kumar, Packer & Koller, 2010) extends the paradigm of curriculum learning (Bengio, Louradour, Collobert & Weston, 2009), where a model is learned by adding gradually more difficult instances during training. The standard SPL model assigns a weight to each instance in the learning objective and adds a self-paced regularizer that determines these weights for a given learning pace. So far, only few works have explored self-paced learning for segmentation. Wang *et al.* (Wang *et al.*, 2018) presented an SPL method for lung nodule segmentation where the weight of each 3D image in the loss is controlled by the SPL regularizer. Recently, Ma *et al.* (Ma *et al.*, 2020) proposed a first self-paced approach for multi-view co-training. Whereas standard co-training techniques adopt a "draw without replacement strategy" which may lead to learning incorrect predictions, this approach adds co-regularization terms in the loss function to select pseudo-labeled instances dynamically during training. Our method, which extends the standard JSD agreement loss to consider prediction uncertainty, significantly differs from this approach based on cross-view correlation. Moreover, while the approach in (Ma *et al.*, 2020) requires an alternating optimization scheme, where pseudo-labels are updated separately from network parameters, our self-paced co-training model can be trained in an end-to-end manner without explicitly computing pseudo-labels.

2.3 The proposed method

An overview of the proposed method for semi-supervised segmentation is shown in Figure 2.1. Let $\mathcal{S} = \{(\mathbf{x}_s, \mathbf{y}_s)\}_{s=1}^{|\mathcal{S}|}$ be a small set of labeled examples, where each $\mathbf{x}_s \in \mathbb{R}^{|\Omega|}$ is an image and $\mathbf{y}_s \in \{0,1\}^{|\Omega|\times|C|}$ is the corresponding ground-truth segmentation mask. Here, $\Omega \subset \mathbb{Z}^2$ denotes the set of image pixels and C the set of segmentation classes. Given labeled images $\mathcal{S}$ and a larger set of unlabeled images $\mathcal{U} = \{\mathbf{x}_u\}_{u=1}^{|\mathcal{U}|}$, the proposed co-training method learns K segmentation networks corresponding to the different views. Each network f^k is parameterized by a set of weights θ^k.

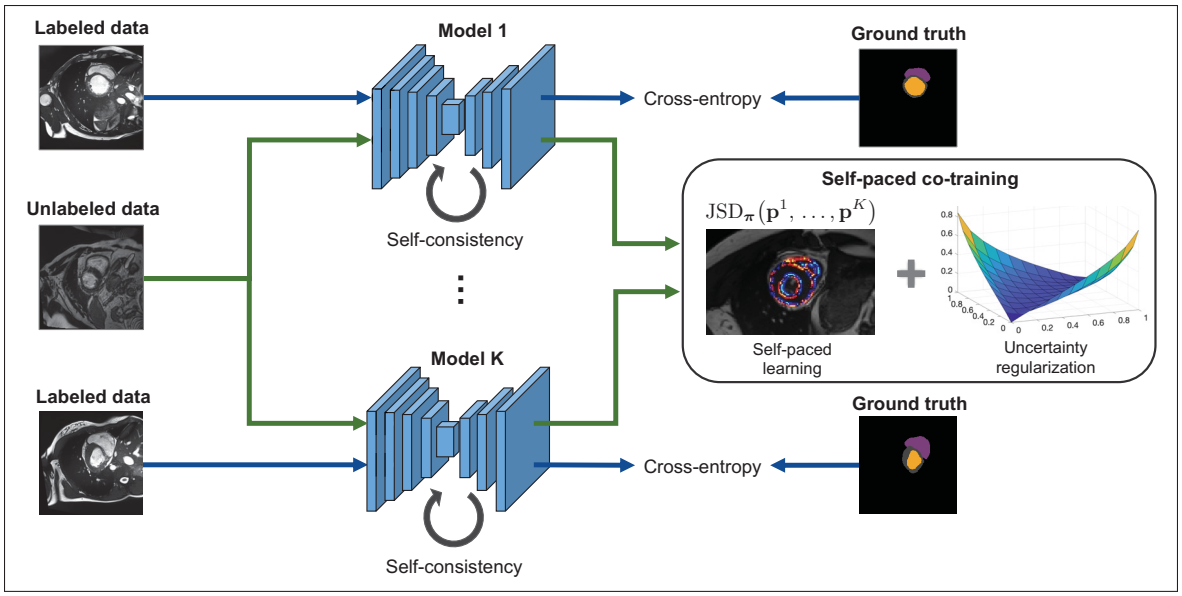


Figure 2.1 Diagram of self-paced and self-consistent co-training for semi-supervised segmentation. The proposed method consists of three different losses: 1) a pixel-wise supervised loss ℓ_{sup} for labeled images; 2) a self-paced co-training loss ℓ_{spc} encouraging the K segmentation models to agree on increasingly harder regions in unlabeled images; 3) a self-consistency loss ℓ_{reg} based on temporal ensembling that regularizes the prediction of individual models.

Labeled and unlabeled images are exploited jointly during training by minimizing the following loss function:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{sup}}(\theta; \mathcal{S}) + \lambda_1 \mathcal{L}_{\text{SPC}}(\theta; \mathcal{U}) + \lambda_2 \mathcal{L}_{\text{reg}}(\theta; \mathcal{U}). \tag{2.1}$$

The supervised loss $\mathcal{L}_{\text{sup}}(\cdot)$ encourages individual networks to predict segmentation outputs for labeled data that are close to the ground truth:

$$\mathcal{L}_{\text{sup}}(\theta; \mathcal{S}) = \frac{1}{K|\mathcal{S}|} \sum_{k=1}^{K} \sum_{(\mathbf{x}_s, \mathbf{y}_s) \in \mathcal{S}} \ell_{\text{sup}}\big(f^k(\mathbf{x}_s; \theta^k), \mathbf{y}_s\big). \tag{2.2}$$

While any segmentation loss can be considered, in this work, we use the well-know cross-entropy loss for $\ell_{\text{sup}}(\cdot)$. Let $\mathbf{p} = f(\mathbf{x}; \theta) \in [0, 1]^{|\Omega| \times |C|}$ be the class probability map predicted by a

network f, cross-entropy is defined as

$$\ell_{\text{sup}}(\mathbf{p}, \mathbf{y}) \;=\; -\sum_{i\in\Omega}\sum_{j\in C} y_{ij}\log(p_{ij}). \tag{2.3}$$

Two separate loss terms are employed to leverage unlabeled data in the learning process. The first one, $\mathcal{L}_{\text{SPC}}(\cdot)$, implements our self-paced co-training strategy that lets segmentation networks learn from gradually-harder examples over training epochs. This strategy also incorporates an uncertainty regularizer encouraging models to become confident during training. The second loss term, $\mathcal{L}_{\text{reg}}(\cdot)$, enhances learning by applying self-consistency regularization on the models of each view. This regularization technique, based on temporal ensembling, improves the robustness of individual networks when training them with few labels. In the following subsections, we present the proposed self-paced learning and self-consistency regularization co-training losses in greater details.

2.3.1 Self-paced co-training

We propose a self-paced strategy where high-confidence regions of unlabeled images are first considered in the loss, and those with lower confidence are gradually incorporated during training. We define our self-paced co-training task as the following optimization problem:

$$\begin{aligned}
&\min\; \mathcal{L}(\{\theta^k\}, \{\widehat{\mathbf{y}}_{ui}\}, \{\mathbf{w}_{ui}\}) \;= \\
&\frac{1}{|\mathcal{U}|}\sum_{\mathbf{x}_u\in\mathcal{U}}\sum_{k=1}^{K}\sum_{i\in\Omega} w_{uik}\, D_{\text{KL}}\big(\mathbf{p}_{ui}^k \,||\, \widehat{\mathbf{y}}_{ui}\big) \;+\; \mathcal{R}_\gamma(w_{uik}) \\
&\text{s.t.}\; \sum_{j\in C}\widehat{y}_{uij} \;=\; 1,\; \forall u, \forall i \\
&\quad \widehat{y}_{uij} \in [0,1],\; \forall u, \forall i, \forall j;\; w_{uik} \in [0,1],\; \forall u, \forall i, \forall k
\end{aligned} \tag{2.4}$$

In this formulation, $\widehat{\mathbf{y}}_{ui}$ is the soft pseudo-label vector of pixel i in unlabeled image $\mathbf{x}_u$, D_{KL} is the Kullback–Leibler (KL) divergence which imposes the prediction $\mathbf{p}_{ui}^k$ of each network f^k to agree with $\widehat{\mathbf{y}}_{ui}$. The importance of pixel i in $\mathbf{x}_u$ on the loss for model f^k is controlled by self-paced learning weight $w_{uik} \in [0, 1]$ that is optimized jointly with $\mathbf{p}_{ui}^k$ and $\widehat{\mathbf{y}}_{ui}$. As in traditional self-paced learning methods (Kumar *et al.*, 2010), $\mathcal{R}$ is a regularization function parameterized by a learning pace parameter $\gamma \geq 0$, such that w_{uik} decreases monotonically with $D_{\mathrm{KL}}\big(\mathbf{p}_u^k \,||\, \widehat{\mathbf{y}}_u\big)$ and increases monotonically with γ. In other words, w_{uik} should be smaller for pixels i of an image $\mathbf{x}_u$ that are more difficult to predict for model f^k (i.e., having a larger loss) and should increase when using a larger learning pace γ.

Choice of self-paced regularization A common choice for regularization term $\mathcal{R}$ is a simple linear function: $\mathcal{R}_\gamma(w_{uik}) = -\gamma\, w_{uik}$. This choice leads to a binary solution where $w_{uik} = 1$ if $D_{\mathrm{KL}}\big(\mathbf{p}_u^k \,||\, \widehat{\mathbf{y}}_u\big) \leq \gamma$ and $w_{uik} = 0$ otherwise (Kumar *et al.*, 2010). In our setting, this simple solution has two considerable drawbacks. First, since the weights are binary, all selected pixels (i.e., $w_{uik} = 1$) for a given learning pace γ have an equal importance in the loss. This is similar to a hard attention mechanism, which typically performs worse than soft-attention for vision tasks (Fu *et al.*, 2019). More importantly, if γ is set too small, very few pixels will contribute to the loss, thereby impeding the learning. To alleviate these problems, we instead employ the following quadratic regularization function: $\mathcal{R}_\gamma(w_{uik}) = \gamma\big(\frac{1}{2}w_{uik}^2 - w_{uik}\big)$.

Self-paced co-training loss and weights update As in standard self-paced learning techniques, the learning weights and model parameters are optimized separately in an alternating manner. However, unlike the approach in (Kumar *et al.*, 2010), our method does not require the explicit computation of pseudo-labels. Learning weights are updated according to the following theorem.

Theorem 1. *Given fixed model parameters* $\{\theta^k\}$ *and pseudo-labels* $\{\widehat{\mathbf{y}}_{ui}\}$*, the optimal learning weights can be obtained as*

$$w_{uik}^* \;=\; \max\left(1 - \frac{1}{\gamma} D_{\mathrm{KL}}\big(\mathbf{p}_{ui}^k \,||\, \widehat{\mathbf{y}}_{ui}\big),\; 0\right) \tag{2.5}$$

The proof of Theorem 1 is shown in Appendix I.1.

In practice, we replace the 0 in Eq. (2.5) by a small $\epsilon > 0$ to avoid zero-divison in subsequent steps. We see that the self-paced regularization term acts as a soft attention mechanism where the importance of a pixel on the loss of model f^k is inversely related to the divergence between the prediction of model f^k and the pseudo-label for that pixel.

We show in the next theorem that the problem of finding pseudo-labels and network parameters amounts to minimizing a generalized form of Jensen-Shannon divergence $\mathrm{JSD}_{\boldsymbol{\pi}}\big(\mathbf{p}^1, \ldots, \mathbf{p}^K\big) = \mathcal{H}\big(\sum_k \pi_k \, \mathbf{p}^k\big) - \sum_k \pi_k \, \mathcal{H}\big(\mathbf{p}^k\big)$.

Theorem 2. *For a fixed set of learning weights* $\{\mathbf{w}_{ui}\}$*, learning the model parameters* $\{\theta^k\}$ *for optimal pseudo-labels corresponds to the following problem:*

$$\min_{\{\theta^k\}} \; \frac{1}{|\mathcal{U}|} \sum_{\mathbf{x}_u \in \mathcal{U}} \sum_{i \in \Omega} \rho_{ui} \, \mathrm{JSD}_{\boldsymbol{\pi}_{ui}}\big(\mathbf{p}^1_{ui}, \ldots, \mathbf{p}^K_{ui}\big) \tag{2.6}$$

$$\textit{with} \;\; \pi_{uik} \;=\; \frac{w_{uik}}{\rho_{ui}}; \quad \rho_{ui} \;=\; \sum_{k=1}^{K} w_{uik}$$

The proof of Theorem 2 is shown in Appendix I.2.

Intuitively, the formulation in Eq. (2.6) imposes individual networks to give, for each unlabeled image pixel, a prediction similar to the confidence-weighted average of all models. Additionally, the importance of each pixel in the total loss is weighted by coefficient ρ_{ui} that corresponds to the total confidence of models for this pixel.

Setting the learning pace parameter A common challenge in self-paced learning methods is finding a suitable value for the learning pace parameter γ: an overly small γ will select too few pixels for co-training, which impedes learning, whereas a too large γ will ignore the relative difficulty of individual pixels and amounts to having no self-paced learning. To find a good range of values for γ, we consider the optimal solution for the self-paced weights w_{uik} in Eq. (2.5) and for pseudo-labels $\widehat{\mathbf{y}}_{ui}$ in Eq. (A I-6). Combining both, we have that pixel i of image $\mathbf{x}_u$ is selected for model f^k (i.e., non-zero weight w_{uik}) if $D_{\mathrm{KL}}\big(\mathbf{p}^k_{ui} \,||\, \sum_{k'} \pi_{uik} \, \mathbf{p}^{k'}_{ui}\big) \leq \gamma$. As it is a divergence, the left-side term of this inequality has a lower-bound of 0. Following (Lin, 1991), it

can also be shown that this term is upper-bounded by $-\log_2(\pi_{uik})$. Furthermore, if we use a small ϵ instead of zero when updating the weights in (2.5), we have that $w_{uik} \in [\epsilon, 1]$ and thus $-\log_2(\epsilon/K) = \log_2(K/\epsilon)$ is also an upper bound. Summing up, γ can be increased following the $[0, \log_2(K/\epsilon)]$ range.

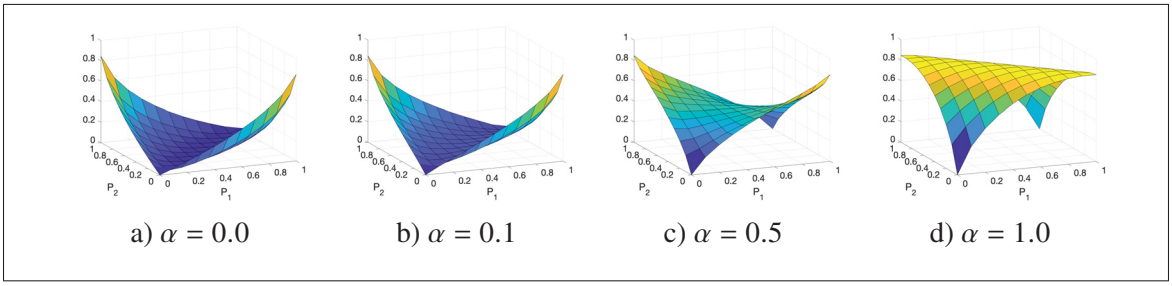


Figure 2.2 Illustration of the proposed entropy regularized JSD between two Bernoulli distributions P_1 and P_2, for different α values. When using $\alpha = 0$, we have the standard JSD which is zero when $P_1 = P_2$ regardless of the confidence (i.e., entropy). As α is increased toward 1, the loss encourages both the agreement and confidence of distributions.

2.3.2 Uncertainty regularization

As defined in Eq. (2.6), our self-paced co-training method, based on generalized JSD, extends the traditional multi-view learning approach where inter-model agreement is typically measured with the standard JSD. An important drawback of JSD is that it enforces models to agree with each other, but does not require them to be confident in their prediction. This is illustrated in Figure 2.2 (left), where we show the JSD between two Bernoulli distributions P_1 and P_2 (i.e., P is the probability of class 1, and $1-P$ the probability of class 2). As can be seen, a JSD of 0 is obtained when $P_1 = P_2$, whether the predictions are confident (e.g., $P_1 = P_2 = 1$) or not (e.g., $P_1 = P_2 = 0.5$).

A powerful technique for semi-supervised learning, called entropy minimization (Grandvalet & Bengio, 2004), consists in increasing the confidence of predictions for unlabeled examples. Based on this idea, we propose an entropy regularizer over JSD that encourages models to agree while also making them confident. This regularized divergence, which is

parameterized by $\alpha \in [0, 1]$, is defined as follows:

$$\mathrm{JSD}^{\alpha}(\mathbf{p}^1, \ldots, \mathbf{p}^K) \;=\; \mathcal{H}\Big(\frac{1}{K}\sum_{k=1}^{K}\mathbf{p}^k\Big) \;-\; \frac{(1-\alpha)}{K}\sum_{k=1}^{K}\mathcal{H}\big(\mathbf{p}^k\big). \tag{2.7}$$

When $\alpha = 0$, we get the standard definition of JSD. On the other hand, for $\alpha = 1$, JSD^{α} simply measures the entropy of the mean prediction which is 0 only when 1) all models are 100% confident and 2) the models agree with each other. Since entropy is non-negative, we have that $\mathrm{JSD}^{\alpha} \geq \mathrm{JSD}^{\alpha'}$ for any $\alpha \geq \alpha'$. As shown in Figure 2.2, while using $\alpha > 0$ increases the confidence of predictions, it also makes the function non-convex (standard JSD is convex). This may pose a problem in initial stages of training, since incorrect predictions are forced to become more confident until the optimization gets stuck in a poor local minimum. In practice, we avoid this issue by using the convex JSD at the beginning of training and then increasing α slowly to reach a fixed value.

Incorporating the proposed entropy regularization in our self-paced co-training method, we finally get the following loss:

$$\mathcal{L}_{\mathrm{SPC}}(\theta; \mathcal{U}) \;=\; \frac{1}{|\mathcal{U}|}\sum_{\mathbf{x}_u \in \mathcal{U}}\sum_{i \in \Omega}\rho_{ui}\,\mathrm{JSD}^{\alpha}_{\boldsymbol{\pi}_{ui}}\big(\mathbf{p}^1_{ui}, \ldots, \mathbf{p}^K_{ui}\big) \tag{2.8}$$

where ρ_{ui} and $\boldsymbol{\pi}_{ui}$ are defined as in Eq. (2.6), and

$$\mathrm{JSD}^{\alpha}_{\boldsymbol{\pi}}\big(\mathbf{p}^1, \ldots, \mathbf{p}^K\big) \;=\; \mathcal{H}\Big(\sum_{k=1}^{K}\pi_k\,\mathbf{p}^k\Big) \;-\; \big(1-\alpha\big)\sum_{k=1}^{K}\pi_k\,\mathcal{H}\big(\mathbf{p}^k\big). \tag{2.9}$$

2.3.3 Self-consistent co-training

While co-training promotes consistency across different models, it has also been shown that imposing consistency between the predictions of a single model at different training iterations can also improve its robustness in a semi-supervised setting (Antti & Valpola, 2017). Based on this idea, we propose to incorporate a self-ensembling strategy in our co-training method,

where we replace each segmentation model f^k by two models: a Teacher f_T^k and a Student model f_S^k. The two models share the same architecture, however the Teacher's parameters are a temporal ensembling of the student's across different training steps. Specifically, we update the Teacher's parameters at step t, denoted as $\theta_T^{(t)}$, using an exponential moving average of the Student's parameters $\theta_S^{(t)}$:

$$\theta_T^{(t)} = \beta\theta_T^{(t-1)} + (1-\beta)\theta_S^{(t)}. \tag{2.10}$$

The Teacher closely follows the Student for $\beta \approx 0$, whereas it has a longer-term memory when β is near 1. In the latter case, each step the Student takes contributes slightly to the Teacher and, therefore, the Teacher is likely to have a smoother learning. Following the literature (Antti & Valpola, 2017), we adopt a β of 0.99.

We impose a self-consistency loss that minimizes the L_2 distance between the predictions of Teacher-Student pairs for unlabeled images $\mathbf{x}_u \in \mathcal{U}$ under random geometric transformations τ:

$$\mathcal{L}_{\text{reg}}(\theta;\mathcal{U}) = \frac{1}{K\,|\mathcal{U}|}\sum_{k=1}^{K}\sum_{\mathbf{x}_u\in\mathcal{U}}\left\|\tau\left(f_T^k(\mathbf{x}_u)\right) - f_S^k\left(\tau(\mathbf{x}_u)\right)\right\|^2. \tag{2.11}$$

Here, the same transformation τ is applied to the Teacher's prediction so that it aligns with the Student's output for the transformed input image. In this work, we considered random rotations for τ. Moreover, while KL divergence can also be employed to measure prediction agreement, like most temporal-ensembling approaches, we used L_2 distance as it leads to a smoother optimization. Unlike L_2 which has a bounded gradient, the gradient of KL may become large, especially in initial training iterations. Thus, it can overpower the gradient of the supervised loss and cause the learning to fail. The overall optimization process, combining all three loss terms, is summarized in Algorithm 2.1.

2.4 Experiments

To evaluate the performance of the proposed self-paced and self-consistent co-training method, we test it on three medical image segmentation tasks involving MRI and CT data. In the next

Algorithm 2.1 Training of the self-paced and self-consistent co-training model.

Input: Labeled dataset $\mathcal{S}$, unlabeled dataset $\mathcal{U}$, and number of co-training models K
Output: Model parameters $\{\theta^k\}$

Randomly initialize network parameters θ^k, $\forall k$;
Initialize learning pace: $\gamma \leftarrow \gamma_0$;
for epoch = 1, ..., n_{epochs} **do**
  for n = 1, ..., n_{iter} **do**
    Sample training batch $\{\mathcal{S}_n, \mathcal{U}_n\}$;
    For all $\mathbf{x}_u \in \mathcal{U}_n$, compute learning weights w_{uik} using Eq. (2.5);
    According to Eq. (2.1), do a batch gradient descent step on the Student models' parameters θ_S^k;
    Update the Teacher models' parameters θ_T^k based on Eq. (2.10);
  end for
  Adjust the SGD learning rate;
  Update learning pace: $\gamma \leftarrow \gamma \times$ increaseFactor;
end for
return $\{\theta^k\}$;

subsections, we describe the datasets and performance metrics employed in the evaluation, the experimental setup, and implementation details of our method.

2.4.1 Datasets and metrics

Datasets

Our experiments are conducted on three clinically-relevant benchmark datasets: Automated Cardiac Diagnosis Challenge (ACDC) (Bernard *et al.*, 2018), Spleen sub-task dataset of the Medical Segmentation Decathlon Challenge (Simpson, Michela & *et al.*, 2019), and the Prostate MR Image Segmentation (PROMISE) 2012 Challenge dataset (Litjens *et al.*, 2014).

ACDC: This dataset consists of 200 MRI scans from 100 patients, including 20 healthy patients, 20 patients with previous myocardial infarction, 20 patients with dilated cardiomyopathy, 20 patients with an hypertrophic cardiomyopathy, and 20 patients with abnormal right ventricle. Scans correspond to end-diastolic (ED) and end-systolic (ES) phases, and were acquired on

1.5T and 3T systems with resolutions ranging from 0.70×0.70 mm to 1.92×1.92 mm in-plane and 5 mm to 10 mm through-plane. Three cardiac regions are labeled in the ground-truth: left ventricle (LV), right ventricle (RV) and myocardium (Myo). In our experiments, we treat the extraction of each region as a separate binary segmentation task. For our experiments, we used a split of 75 subjects (150 scans) for training and 25 subjects (50 scans) for testing. Slices within 3D-MRI scans were considered as 2D images, themselves randomly cropped into patches of size 192×192. These patches are fed as input to the network.

Spleen: This dataset consists of total 61 CT scans (only 41 were given with ground truth). For our experiments, 2D images are obtained by slicing the high-resolution CT volumes, followed by a max-min normalization to the $[0, 1]$ range. Each slice is then resized to 256×256. We split the dataset into training and testing sets, comprised respectively of 36 and 5 CT scans.

Prostate: This third dataset, which is provided by Radboud University, consists of 50 3D T2-weighted MRIs of the prostate region with expert annotations for two structures: peripheral zone and central gland. This dataset was split into training and testing sets containing the MRIs of 40 and 10 patients, respectively. Once again, slices within MRIs were treated as 2D images that are randomly cropped into input patches of size 192×192.

Metrics

Two metrics are used to evaluate the segmentation accuracy of tested methods: Dice similarity coefficient (DSC) and Hausdorff distance (HD). DSC measures the degree of overlap between the segmentation region and ground truth, and is defined as

$$\mathrm{DSC}(G, S) = \frac{2|S \cap G|}{|S| + |G|}, \tag{2.12}$$

where S is the predicted labels and G is the corresponding ground truth labels. DSC values range from 0 to 1, a higher value representing a better segmentation.

Haussdorf distance (HD) is a boundary distance metric which measures the largest distance between a point in S and its nearest point in G. A smaller HD value indicates a better

segmentation. HD is defined as

$$\mathrm{HD}(G, S) \;=\; \max\left\{d(G, S),\; d(S, G)\right\} \tag{2.13}$$

where $d(S, G)$ is the maximum of nearest-neighbor distances from S to G. In our results, we report the HD in millimeters.

2.4.2 Experimental setup

We first use different levels of supervision to assess the stability of our proposed method. For ACDC and Prostate, we tested the following two settings: 10% and 5% of the training set as labeled data. Since the Spleen dataset is more challenging to segment, considering only 5% of training examples as labeled leads to the collapse of tested algorithms. To avoid this problem, we instead used labeled data ratios of 10% and 7% for this dataset. Moreover, to measure the impact of our self-paced learning and self-consistency losses, we performed an ablation study where we disable one of these losses while keeping the other. Finally, to investigate our method's performance and scalability in a multi-view setting, we tested it with more than two segmentation models.

We compare our method against several baselines. To have upper and lower bounds on performance, we first include full-supervision and semi-supervision baselines. The full-supervision baseline considers all training examples as labeled. On the other hand, the semi-supervision baseline uses only the partial subset of labeled examples (i.e., 10%, 7% or 5%) and ignores the remaining unlabeled images during training. As our proposed method extends standard co-training, we consider the model with only JSD as another strong baseline called Co-training. Moreover, we compare our method with four state-of-the-art semi-supervised segmentation approaches: Entropy minimization (Grandvalet & Bengio, 2004), Deep adversarial networks (DAN) (Zhang *et al.*, 2017), Mean Teacher (MT) (Antti & Valpola, 2017), and Uncertainty-aware Mean Teacher (UA-MT) (Yu *et al.*, 2019). We keep the same underlying architecture, optimization procedure, and data augmentation across all tested methods. For main

experiments, we adopt a dual-view setting for standard co-training and our method. As in (Peng *et al.*, 2020a), we use soft-voting to aggregate the prediction of individual models into the final segmentation.

For a fair comparison, we implemented all tested methods by ourselves in a single framework, using the same segmentation backbone network and optimization strategy, and selected the hyper-parameters of all methods based on grid search. For Mean Teacher, which involves two models and a data transformation procedure, we applied data transformations on input images for the Student model and on output predictions for the Teacher. We update the Teacher using Eq. (2.10) and, following (Peng *et al.*, 2020a), use this model's output as the segmentation prediction to measure performance.

2.4.3 Implementation details

Although any 2D segmentation network can be used, we employed the popular light-weight architecture ENet (Adam *et al.*, 2016) as the underlying segmentation network, as it offers a good trade-off between accuracy and speed. This architecture, which was developed for efficient semantic segmentation, contains about 84 times less trainable parameters than the well-known UNet architecture (0.37 M compared to 31.04 M for UNet). It employs a convolutional block with short skip connections, called bottleneck block, and is comprised of different 7 stages: an initial stage of regular convolutions, followed by 5 stages with different numbers of bottleneck blocks, and a final stage of 1×1 convolutions to generate the final segmentation probability map.

All experiments were carried out using the same training setting with the ENet architecture, rectified Adam optimizer and learning rate warm-up strategy based on cosine decay (Loshchilov & Hutter, 2016). Training was performed for a total of 100 epochs, each one including 200 update iterations. The learning rate was initialized to 1×10^{-3} for ACDC, 1×10^{-4} for Prostate and 3×10^{-5} for Spleen. For all datasets, this learning rate was increased by a factor of 300 in the first 10 epochs and then decreased with a cosine scheduler for the following 90 epochs.

We used standard data augmentation techniques on-the-fly to avoid over-fitting, including randomly cropping and rotation. For a fair comparison, we keep data augmentation the same across different methods for each segmentation task. For our proposed method, the total loss consists of three terms: supervision loss $\mathcal{L}_{\text{sup}}$ for labeled data, self-paced co-training loss $\mathcal{L}_{\text{SPC}}$ and self-consistency loss $\mathcal{L}_{\text{reg}}$ for unlabeled data. As defined in Eq. (2.1), these loss terms are balanced with two hyper-parameters λ_1 and λ_2. Based on grid search, these coefficients were set as follows: $\lambda_1 = 0.5$, $\lambda_2 = 4$ for ACDC, $\lambda_1 = 0.1$, $\lambda_2 = 4$ for Prostate, and $\lambda_1 = 0.5$, $\lambda_2 = 12$ for Spleen. For all three datasets, α was initialized to 0 and then gradually increased to a value near $\alpha = 10^{-4}$ over training. Using a larger value for α led to a worse performance. We believe this is due to the fact that entropy is a concave function with exponentially growing gradient near values of 0 and 1. A small $\alpha = 10^{-4}$ is therefore necessary to avoid the gradient from this term dominating the learning. A small coefficient was also used in previous approaches based on entropy regularization (Grandvalet & Bengio, 2004).

The learning pace parameter γ was set based on the strategy presented in Section 2.3.1. To ensure that the prediction for some pixels is used in the loss of Eq. (2.6), i.e., that not all learning weights are 0, we set the initial pace γ_0 as follows: $\gamma_0 = 0.2$ for ACDC and Spleen with 10% labeled data, $\gamma_0 = 0.4$ for Prostate with 10% and Spleen datasets with 7% labeled data, and $\gamma_0 = 0.5$ for Prostate datasets with 5% labeled data. In all cases, γ was updated so that its upper bound is reached after 50 epochs (see Section 2.3.1 for details). For all experiments except the multi-view analysis, we trained our method and standard co-training using $K = 2$ segmentation networks. In the multi-view analysis, we compare this setting with using $K = 3$ views. The same hyper-parameters as in previous experiments are used in both cases.

To have a fair comparison, the same grid search strategy was used to select the hyper-parameters of all tested approaches. For each method, we report results obtained for the best combination of hyper-parameters found during grid search.

2.5 Results

2.5.1 Comparison to the state-of-art

We first compare our self-paced and self-consistent co-training method against the baselines and state-of-the-art approaches for semi-supervised segmentation (i.e., Entropy minimization, Deep adversarial networks, Mean Teacher, Uncertainty-aware Mean Teacher, and Co-training).

Tables 2.1 and 2.2 give the mean DSC and HD obtained by the tested methods for the ACDC dataset. Reported values are averages (standard deviation in parentheses) over 3 runs with different random seeds. As can be seen, all semi-supervised methods yield improvements compared to training without unlabeled images (i.e., partial supervision). Entropy minimization improves the mean segmentation performance by about 0.4% in terms of DSC compared to the partial supervision baseline, in both the 10% and 5% labeled data settings. This confirms the benefit of making the prediction of a single model more confident. Likewise, standard Co-training outperforms partial supervision by 2.2% for 10% of labeled data, showing that the collaborative training of two models can improve their individual performance. Mean Teacher, which implements temporal ensembling, achieves a mean DSC gain of 3.3% for the same labeled data setting. UA-MT further improves this score by 0.3%, demonstrating the usefulness considering uncertainty in the Mean Teacher framework. However, the best performance on ACDC is obtained by our method with a mean DSC of 87.78% and 86.42%, respectively for 10% and 5% labeled examples. This DSC performance is only 2.8% and 4.1% less than full-supervision. Considering segmentation classes separately, we observe that the highest improvements by our method are for RV (+6.3%) and Myo (+4.9%), which are the most difficult regions to segment. The proposed method also leads to an important reduction of HD for all classes and labeled data ratios. Using 5% of labeled data, our method decreases the mean HD substantially from 17.5 mm to 7.2 mm compared to the partial supervision baseline. In comparison, the strong UA-MT baseline yields a much higher mean HD of 12.3 mm for the same setting. This shows our method's greater ability to regularize segmentation boundaries and avoid large gaps to the ground-truth region.

Table 2.1 Mean DSC (%) of tested methods on the ACDC dataset, for different ratios of labeled training examples. For our method and Co-training, *avg* is the average performance of the two separate views and *voting* the performance of combining their prediction through voting. Bold font values indicate the best performing method for each labeled data setting. Values are underlined if the improvement over all other approaches is statistically significant ($p < 0.05$) – Ours (voting / avg, resp.) is compared against Co-training (voting / avg, resp.) and all other methods.

Labeled %	Method		ACDC			
			RV	Myo	LV	Mean
100 %	Baseline		89.29 (0.37)	88.30 (0.15)	94.10 (0.32)	90.56 (0.28)
10 %	Baseline		77.51 (0.87)	81.56 (0.46)	91.72 (0.24)	83.60 (0.53)
	Entropy min		78.19 (2.29)	82.04 (0.40)	91.84 (0.27)	84.02 (0.99)
	DAN		82.81 (0.15)	83.77 (0.09)	91.90 (0.25)	86.16 (0.16)
	MT		82.91 (1.55)	85.35 (0.57)	92.37 (0.32)	86.88 (0.81)
	UA-MT		83.63 (0.89)	85.78 (0.16)	92.13 (0.56)	87.18 (0.54)
	Co-training	avg	80.69 (0.59)	83.07 (0.17)	92.40 (0.23)	85.38 (0.33)
		voting	80.88 (0.53)	83.69 (0.32)	92.84 (0.31)	85.80 (0.39)
	Ours	avg	83.55 (0.62)	86.18 (0.37)	92.75 (0.48)	87.50 (0.49)
		voting	**83.85 (0.51)**	**86.42 (0.29)**	**93.06 (0.07)**	**87.78 (0.29)**
5 %	Baseline		72.32 (2.47)	75.69 (0.69)	86.87 (0.79)	78.29 (1.31)
	Entropy min		73.56 (0.51)	75.47 (0.84)	87.25 (1.19)	78.76 (0.85)
	DAN		81.54 (0.53)	80.13 (0.45)	90.96 (0.33)	84.21 (0.44)
	MT		79.52 (0.98)	83.08 (0.60)	91.38 (0.14)	84.66 (0.57)
	UA-MT		81.71 (0.62)	83.08 (0.23)	90.69 (0.83)	85.16 (0.56)
	Co-training	avg	74.96 (1.08)	79.25 (0.04)	90.66 (0.03)	81.62 (0.38)
		voting	75.37 (1.35)	79.41 (0.41)	90.72 (0.05)	81.83 (0.60)
	Ours	avg	**82.35 (0.64)**	**83.82 (0.39)**	91.76 (0.55)	85.98 (0.53)
		voting	**82.33 (0.16)**	**84.46 (0.22)**	**92.47 (0.14)**	**86.42 (0.17)**

The performance of our self-paced and self-consistent co-training method is further demonstrated for the Prostate and Spleen datasets in Tables 2.3 and 2.4. For Prostate, our method achieves significant DSC improvements of 1.2% when using 5% of labeled data, compared to the second-best approach (i.e., UA-MT). An even greater performance boost is obtained on the challenging Spleen dataset, with DSC gains of 2.6% and 2.1% compared to UA-MT for labeled data ratios of 10% and 7%, respectively. Similar improvements are observed in terms of HD for all labeled data ratios. Surprisingly, Co-training obtains a lower mean Dice than the partial supervision baseline for these two datasets. This is due to the fact that the final prediction of

Table 2.2 Mean Hausdorff distance (HD) of tested methods on the ACDC dataset, for different ratios of labeled training examples. For our method and Co-training, *avg* is the average performance of the two separate views and *voting* the performance of combining their prediction through voting. Bold font and underlined values are defined as in Table 2.1.

Labeled %	Method		ACDC			
			RV	Myo	LV	Mean
100 %	Baseline		6.05 (0.49)	3.99 (0.07)	2.87 (0.08)	4.30 (0.21)
10 %	Baseline		15.10 (2.13)	11.00 (2.84)	6.20 (1.29)	10.77 (2.09)
	Entropy min		16.85 (2.12)	8.03 (0.48)	5.77 (0.82)	10.22 (1.41)
	DAN		14.00 (0.11)	6.28 (0.82)	4.46 (1.21)	8.25 (0.71)
	MT		12.30 (0.02)	6.97 (1.00)	4.89 (1.11)	8.05 (0.71)
	UA-MT		17.12 (6.36)	8.02 (0.45)	6.92 (0.38)	10.69 (2.40)
	Co-training	avg	11.86 (1.33)	6.77 (0.27)	5.17 (0.04)	7.93 (0.55)
		voting	**9.17 (0.60)**	6.61 (0.35)	5.01 (0.49)	6.93 (0.48)
	Ours	avg	11.05 (1.33)	**5.20 (0.57)**	3.86 (0.50)	6.67 (0.80)
		voting	10.10 (0.68)	**5.43 (0.40)**	**3.81 (0.58)**	**6.45 (0.56)**
5 %	Baseline		23.74 (5.97)	17.27 (3.39)	11.48 (3.78)	17.50 (4.38)
	Entropy min		24.45 (1.50)	17.30 (2.76)	12.89 (3.60)	18.21 (2.62)
	DAN		12.38 (0.43)	8.14 (1.32)	7.04 (1.54)	9.19 (1.10)
	MT		14.13 (1.93)	12.55 (7.28)	11.02 (1.65)	12.57 (3.62)
	UA-MT		14.85 (1.27)	11.98 (1.94)	10.06 (2.35)	12.30 (1.85)
	Co-training	avg	12.83 (1.22)	7.63 (0.95)	5.27 (0.42)	8.58 (0.86)
		voting	11.07 (0.83)	7.00 (0.24)	5.25 (0.11)	7.77 (0.39)
	Ours	avg	11.96 (0.89)	6.77 (1.52)	5.68 (1.94)	8.14 (1.45)
		voting	**10.22 (0.43)**	**6.30 (0.49)**	**4.93 (0.46)**	**7.15 (0.46)**

this method is the average of two separate networks. For challenging segmentation datasets like Prostate and Spleen, and when having few labeled images, it can happen that one of the two networks gives considerably worse predictions than the other. In this case, the poorly-performing network will hurt the co-training of both models, and the average of predictions will be worse than the prediction of a single network (e.g., Baseline). By exchanging only confident predictions in a self-paced manner, our method can effectively avoid this issue.

To determine whether the improvements achieved by our method are significant, we ran a one-sided paired t-test for each segmentation task and performance metric. In Tables 2.1-2.4, we underline the score of the best-performing method if it is significantly better (i.e., higher for

Table 2.3 Mean DSC and HD of tested methods on the Prostate dataset, for different ratios of labeled training examples. For our method and Co-training, *avg* is the average performance of the two separate views and *voting* the performance of combining their prediction through voting. Bold font and underlined values are defined as in Table 2.1.

Labeled %	**Method**		**Prostate**	
			DSC (%)	**HD (mm)**
100 %	Baseline		87.99 (0.20)	5.04 (0.42)
10 %	Baseline		63.77 (1.76)	9.55 (0.80)
	Entropy min		65.30 (3.43)	11.11 (4.57)
	DAN		72.62 (1.53)	10.73 (1.44)
	MT		75.27 (0.72)	9.92 (1.20)
	UA-MT		75.92 (2.77)	7.22 (0.23)
	Co-training	avg	65.06 (0.60)	8.83 (0.62)
		voting	65.09 (0.67)	7.83 (0.32)
	Ours	avg	74.34 (2.44)	7.03 (0.42)
		voting	**<u>76.16</u> (0.62)**	**<u>6.32</u> (0.54)**
5 %	Baseline		49.97 (0.83)	11.65 (5.25)
	Entropy min		50.21 (1.93)	13.52 (3.95)
	DAN		61.17 (2.43)	25.51 (4.55)
	MT		67.97 (1.88)	11.72 (0.76)
	UA-MT		68.91 (1.76)	11.09 (0.45)
	Co-training	avg	49.20 (3.03)	11.59 (1.03)
		voting	48.92 (2.74)	9.72 (0.92)
	Ours	avg	**<u>70.36</u> (0.14)**	**<u>7.90</u> (1.20)**
		voting	**<u>70.15</u> (0.27)**	**<u>8.18</u> (0.63)**

DSC, smaller for HD) than the second-best one for the same setting. Significance is established when $p < 0.05$. It can be seen that our method yields significant improvements in all but 5 of the 20 test cases (i.e., 5 segmentation tasks $\times$ 2 labeled data ratios $\times$ 2 performance metrics).

2.5.2 Visualization of results

We also confirm the effectiveness of our method by visually comparing segmentation results of tested approaches. Figure 2.3 shows examples of results for test images in the three datasets, when training with 10% labeled data. It can be seen that our method gives smoother segmentation

Table 2.4 Mean DSC and HD of tested methods on the Spleen dataset, for different ratios of labeled training examples. For our method and Co-training, *avg* is the average performance of the two separate views and *voting* the performance of combining their prediction through voting. Bold font and underlined values are defined as in Table 2.1.

Labeled %	**Method**		**Spleen**	
			DSC (%)	**HD (mm)**
100 %	Baseline		94.02 (0.41)	9.40 (4.08)
10 %	Baseline		61.67 (4.15)	84.08 (4.84)
	Entropy min		62.50 (1.13)	71.95 (6.89)
	DAN		71.64 (1.15)	119.43 (9.71)
	MT		86.95 (0.95)	74.45 (13.33)
	UA-MT		85.53 (1.44)	59.88 (3.63)
	Co-training	avg	59.64 (2.56)	84.03 (5.16)
		voting	58.95 (2.86)	73.43 (7.25)
	Ours	avg	**<u>88.35</u> (1.56)**	**<u>43.29</u> (14.66)**
		voting	**<u>88.10</u> (1.57)**	**<u>30.52</u> (5.23)**
7 %	Baseline		58.84 (3.36)	98.92 (2.43)
	Entropy min		61.29 (0.10)	103.09 (4.01)
	DAN		64.84 (3.02)	107.54 (14.43)
	MT		81.32 (3.01)	65.53 (43.63)
	UA-MT		83.37 (4.51)	66.54 (3.09)
	Co-training	avg	57.02 (2.43)	94.40 (7.59)
		voting	58.26 (0.65)	85.55 (8.45)
	Ours	avg	**<u>85.20</u> (2.59)**	83.62 (21.64)
		voting	**<u>85.50</u> (3.57)**	**<u>58.77</u> (7.51)**

contours compared to standard Co-training, MT, and UA-MT, and achieves segmentation predictions closer to the ground-truth mask, despite the low contrast of input images.

To visualize the impact of the uncertainty regularizer in the proposed method, Figure 2.4 plots the prediction entropy maps of our method and standard Co-training for a test image in the Prostate dataset, at different training epochs. Compared to Co-training, the proposed uncertainty-regularized method gives a more confident prediction during training, leading to an improved segmentation.

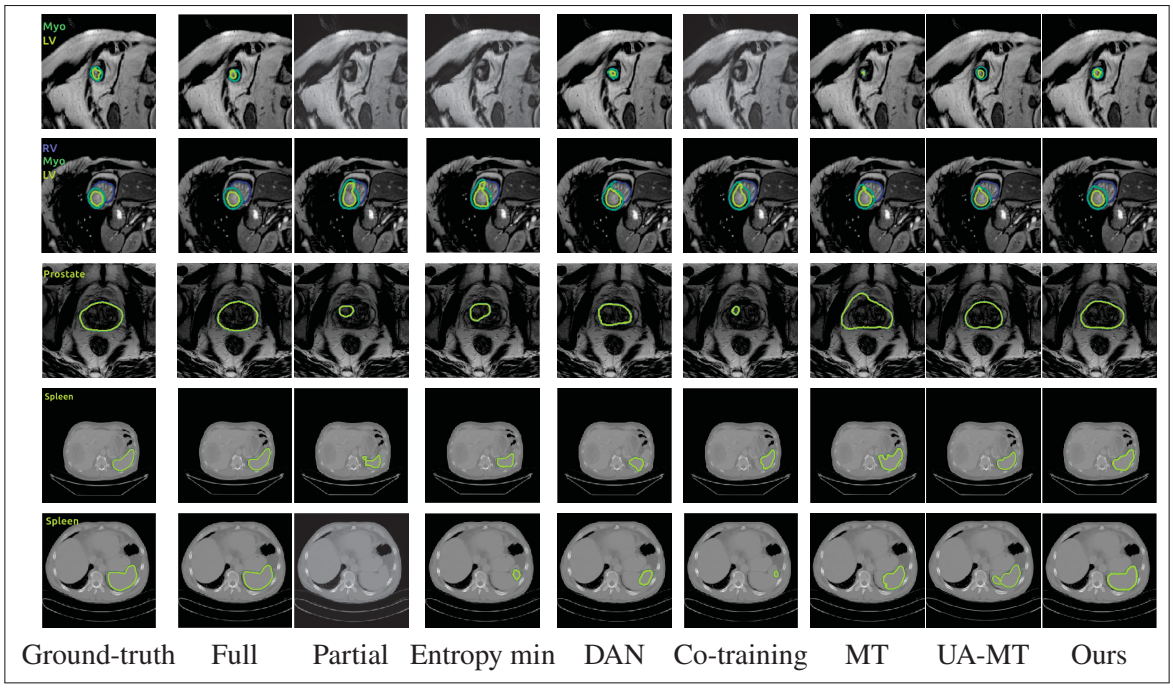


Figure 2.3 Visual comparison of tested methods on test images. **Top two rows**: ACDC dataset. **Middle row**: Prostate dataset. **Bottom two rows**: Spleen dataset. A labeled data ratio of 10% was used for all three datasets. Our method and Co-training were trained in a dual-view setting.

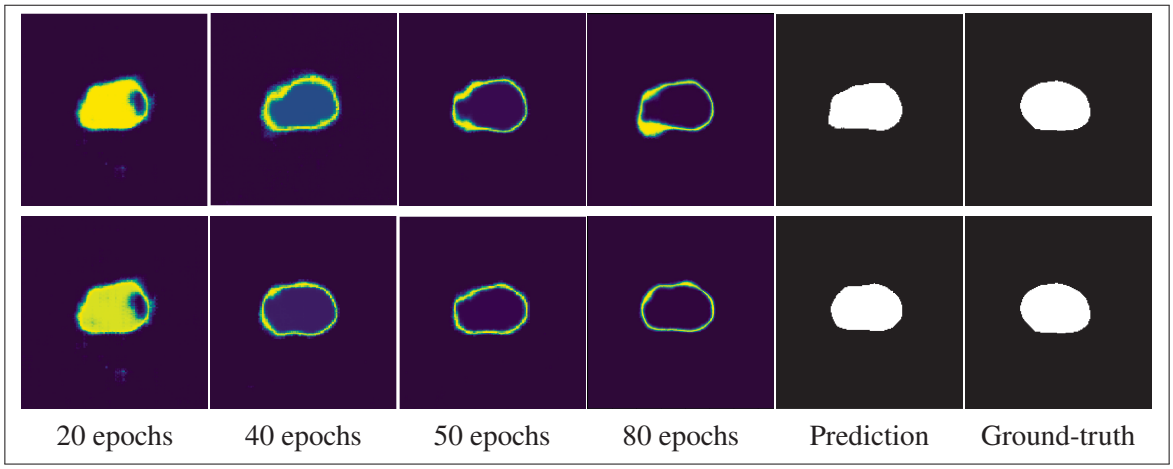


Figure 2.4 Entropy maps, predicted segmentation and ground-truth mask for an image in the Prostate dataset. **Top row**: without our α-entropy JSD loss. **Bottom row**: with the loss. It can be seen that the prediction becomes confident when using the proposed loss.

Table 2.5 Mean DSC (%) of our method with different ablation settings, self-consistency(Self-c) and self-paced learning (Self-pl), on ACDC, Prostate and Spleen, using 5%, 5% and 7% of labeled examples, respectively.

Self-c	Self-pl	ACDC				Prostate	Spleen
		RV	Myo	LV	Mean		
✗	✗	76.28 (1.33)	79.48 (0.61)	90.95 (0.34)	82.24 (0.76)	51.02 (1.62)	58.77 (3.86)
✗	✓	77.87 (0.56)	79.66 (0.02)	91.34 (0.38)	82.96 (0.32)	53.22 (0.83)	61.33 (1.93)
✓	✗	82.05 (0.69)	84.27 (0.59)	92.20 (0.68)	86.17 (0.65)	68.09 (1.91)	83.07 (2.89)
✓	✓	**82.33 (0.16)**	**84.46 (0.22)**	**92.47 (0.14)**	**86.42 (0.17)**	**70.15** (0.27)	**85.50 (3.57)**

Table 2.6 Mean DSC (%) of co-training methods on the ACDC dataset with 10% labeled data, for 2 or 3 views.

Views	Method	ACDC			
		RV	Myo	LV	Mean
2	Co-training	80.88 (0.53)	83.69 (0.32)	92.84 (0.31)	85.80 (0.39)
	Ours	**83.85 (0.51)**	**86.42 (0.29)**	**93.06 (0.07)**	**87.78 (0.29)**
3	Co-training	81.51 (0.19)	84.49 (0.27)	92.84 (0.40)	86.28 (0.29)
	Ours	**83.97** (0.58)	**86.52** (0.42)	**93.22** (0.11)	**87.90 (0.37)**

2.5.3 Ablation study

To assess the respective contribution of the self-paced learning loss ($\mathcal{L}_{\text{SPC}}$) and self-consistency regularization loss ($\mathcal{L}_{\text{reg}}$) in our co-training method, we performed an ablation study where we

Table 2.7 Mean DSC (%) of the baseline, standard co-training and our method on the Spleen dataset with 10% labeled data, when training with images from a single imaging plane (axial, sagittal or coronal) or all planes jointly.

Labeled %	Method	Single view			Multi-view
		Axial	Sagittal	Coronal	
100 %	Baseline	89.74 (0.30)	89.75 (0.19)	89.58 (0.07)	90.14 (0.22)
10%	Baseline	78.28 (0.92)	75.46 (3.22)	74.97 (3.02)	78.69 (2.09)
	Co-training	87.01 (0.78)	86.96 (0.90)	85.48 (0.99)	87.36 (1.22)
	Ours	**88.28 (1.71)**	**88.48 (1.14)**	**88.42 (1.10)**	**89.04 (0.48)**

Table 2.8 Mean DSC (%) of Mean Teacher and our methods on the ACDC dataset with 10% labeled data and different backbone network architectures.

Architectures	Method	ACDC			
		RV	Myo	LV	Mean
Enet	Baseline	77.51 (0.87)	81.56 (0.46)	91.72 (0.24)	83.60 (0.53)
	MT	82.91 (1.55)	85.35 (0.57)	92.37 (0.32)	86.88 (0.81)
	UA-MT	83.63 (0.89)	85.78 (0.16)	92.13 (0.56)	87.18 (0.54)
	Ours	**83.85 (0.51)**	**86.42 (0.29)**	**93.06 (0.07)**	**87.78 (0.29)**
U-Net	Baseline	68.28 (1.61)	79.94 (1.00)	86.41 (0.29)	78.21 (0.89)
	MT	74.62 (1.10)	80.66 (0.61)	86.75 (0.27)	80.68 (0.41)
	UA-MT	74.75 (0.88)	80.92 (0.48)	87.03 (0.43)	80.90 (0.60)
	Ours	**76.04 (0.73)**	**81.87 (0.22)**	**88.17 (0.18)**	**82.03 (0.38)**
SegNet	Baseline	71.82 (0.82)	79.84 (1.19)	89.86 (1.49)	80.51 (1.17)
	MT	80.68 (0.46)	85.17 (0.45)	93.20 (0.14)	86.35 (0.35)
	UA-MT	79.92 (0.22)	85.24 (0.19)	93.20 (0.27)	86.12 (0.23)
	Ours	**82.47 (0.45)**	**86.86 (0.27)**	**93.50 (0.33)**	**87.61 (0.35)**

disable one or both of these losses during training. We carried out this study on the ACDC, Prostate, and Spleen datasets with labeled data ratios 5%, 5%, and 7% respectively, and show the results in Table 2.5.

We see that both the self-paced learning and self-consistency strategies improve performance when used by themselves. Hence, the self-paced learning strategy brings improvements of 0.7%, 2.2% and 2.6% in mean DSC for ACDC, Prostate and Spleen, respectively, while self-consistency alone boosts DSC performance by 3.9%, 17.1% and 24.3% for these datasets. However, combining both strategies results in even greater improvements of 4.2%, 19.1% and 26.7%, demonstrating their synergy and complementary benefit.

2.5.4 Multi-view analysis

In previous experiments, we tested our co-training method in a dual-view setting where two segmentation networks are trained in a collaborative manner. In the next analysis, we assess whether increasing the number of views can further improve segmentation performance. Table 2.6 compares the results of co-training and our method trained with 10% of labeled examples

Table 2.9 Training and inference time of the tested methods, for a batch size of 1.

Method	Views	Training time (ms / batch)	Inference time (ms / batch)
Baseline	1	285	87
Entropy min	1	310	87
DAN	1	450	87
MT	1	380	87
UA-MT	1	395	87
Co-training	2	540	128
	3	760	248
Ours	2	660	128
	3	990	248

from the ACDC dataset, using either 2 or 3 views. We see that jointly training 3 segmentation networks instead of 2 gives a small boost in performance for our proposed method, however this improvement is not statistically significant. Since having more models increases computational requirements, these results suggest that having two views might be best for this segmentation task.

The methods and experiments presented so far generate separate views from the same 2D axial CT/MRI images by applying different randomly-selected transformations to these images. For high-resolution 3D scans, another approach for generating different views is to consider images along different imaging planes of a scan (axial, sagittal and coronal), and train a separate model with images from each plane (Zhou *et al.*, 2019b; Xia *et al.*, 2020a). This is equivalent to considering images as 3D volumes and applying 90 degree rotations around the imaging axes. To evaluate our method in this scenario, we use the Spleen dataset which contains high-resolution CT images and follow the methodology in (Xia *et al.*, 2020a). In this methodology, scans are processed in 3D patches of size $96 \times 96 \times 96$ that are rotated to a different imaging plane for each of the three networks. Asymmetric 3D kernels of size $3 \times 3 \times 1$ are used to ensure that the networks only consider spatial relationship along their respective plane. Table 2.7 reports the results of models trained separately with each view or trained jointly, for the Co-training baseline and our method using 10% of labeled training examples. Comparing individual views

in the semi-supervised setting, we find that axial images give the highest mean DSC for all methods. This is expected since CT scans were acquired along the axial plane, thus 2D images in this plane have the highest resolution. Moreover, we observe a consistent improvement when combining the three views in a co-training framework. Comparing the different approaches, our method outperforms the Partial supervision baseline and standard Co-training both for the single view and multi-view scenarios. Specifically, our method yields mean DSC improvements of 10.35% and 1.68% with respect to these approaches, respectively. All improvements are statistically significant.

2.5.5 Impact of network architecture

To evaluate the robustness of the proposed method to different backbone architectures, we also employed UNet (Ronneberger *et al.*, 2015) and SegNet (Badrinarayanan *et al.*, 2017) as the underlying segmentation network. The U-Net architecture consists of a contracting path and an expansive path. The contracting path follows the architecture of a convolutional network, which consists of the repeated application of two 3×3 convolutions, each followed by a rectified linear unit (ReLU) and a max pooling operation for downsampling. Every step in the expansive path consists of an upsampling of the feature map followed by a convolution that halves the number of feature channels, a concatenation with the correspondingly cropped feature map from the contracting path, and two 3×3 convolutions followed by a ReLU. At the final layer, a 1×1 convolution is used to map each feature vector to the desired number of classes. The number of UNet architecture parameters is 31.04 M. Finally, the SegNet architecture has an encoder and a corresponding decoder, followed by a final pixel-wise classification layer. The encoder consists of 13 convolutional layers identical to the first 13 layers in the VGG16 network. Each encoder layer has a corresponding decoder layer. The final decoder output is fed to a multi-class soft-max classifier to produce class probabilities for each pixel. SegNet has a total of 29.46 M trainable parameters.

The DSC performance of our method, Mean Teacher and UA-MT obtained with different backbone architectures, for the ACDC data with 10% of labeled data, is given in Table 2.8.

Results show our method to provide consistently-higher accuracy than UA-MT for all backbone networks. Comparing the segmentation architectures to each other, we find that ENet and SegNet yield a similar mean DSC, which is about 5.6% greater than UNet. Considering that ENet requires much less computation and memory than SegNet, we conclude that this architecture is best for our method.

Moreover, to evaluate the runtime efficiency of our method, we provide in Table 2.9 the average training time and inference time of tested approaches using ENet as backbone network and a batch size of 1. The baseline model needs to compute only a single loss per pass, thus has the lowest training times. Although Mean Teacher uses two networks in training, only the Student model is updated via back-propagation (the Teacher's parameters are updated with EMA of the Student's). As Mean Teacher, dual-view Co-training also requires training two models. However, the parameters of these models can be updated in parallel, instead of sequentially like Mean Teacher. Because it combines self-ensembling and self-paced co-training, our method requires more computations than Co-training, resulting in a 22% longer training time than this approach in the dual-view setting. In terms of the inference time, both our method and Co-training need to do a forward pass on two separate networks, which increases computations compared to other approaches. As in training, this could also be done in parallel to speed-up inference.

2.6 Discussion and conclusion

We proposed a self-paced and self-consistent co-training method for semi-supervised image segmentation. Our method extends standard co-training by focusing first on easier regions of unlabeled images, and by encouraging both consistency and confidence across the different models during training. Our self-paced learning strategy uses a end-to-end differentiable loss based on generalized JSD to dynamically control the importance of individual pixels on co-training the different segmentation networks. Moreover, a self-consistency loss based on temporal ensembling is used to further regularize the training of individual models and improve performance when annotated data is very limited. We evaluated the potential of our method in three challenging segmentation tasks, including images of different modalities. Experimental

results showed our proposed method to outperform state-of-art approaches for semi-supervised segmentation and yield a performance close to full-supervision while using only a small fraction of the labeled data.

A limitation of the proposed method is the need to run multiple segmentation networks, which increases the computational requirements. Although parallel computation techniques can be adopted to speed up training and inference, this limitation could be addressed alternatively by creating a single model that distillates the knowledge of individual models across views, similar to (Antti & Valpola, 2017). One could also reduce training and inference times by having the co-trained networks share some of their layers, for example allowing only the last few layers of the decoder to differ. Another potential drawback of our method is the need to balance different loss terms that may compete against one another during training. To alleviate this problem, a useful extension of this work could be to investigate self-tuning mechanisms which can adapt more efficiently to new datasets. As future work, we plan to extend our method to the segmentation of 3D and multi-modal images. We will also investigate other strategies for self-paced learning and self-consistency in our co-training framework.

CHAPTER 3

CAT: CONSTRAINED ADVERSARIAL TRAINING FOR ANATOMICALLY-PLAUSIBLE SEMI-SUPERVISED SEGMENTATION

Ping Wang[1] , Jizong Peng[1] , Marco Pedersoli[1] , Yuanfeng Zhou[2] , Caiming Zhang[2] , Christian Desrosiers[1]

[1] Department of Software and IT Engineering, École de Technologie Supérieure, 1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3

[2] School of Software, Shandong University, 1500 Middle of Shunhua Road, Jinan, Shan Dong, China 250101

Article published in Journal «IEEE Transaction on Medical Imaging» in December 2022.

3.1 Introduction

Segmentation is a fundamental problem in medical image analysis. Recent methods for this task, based on deep neural networks, can achieve high performance even when labeled data is limited (Miyato *et al.*, 2019; Grandvalet & Bengio, 2004; Qiao *et al.*, 2018; Antti & Valpola, 2017). However, despite their high accuracy, these methods may still predict segmentations considered anatomically invalid by clinicians (Painchaud *et al.*, 2020). Such predictions can severely impact downstream analyses which rely on anatomical measures, and often require a costly manual step to correct segmentation errors. Incorporating prior knowledge into image segmentation algorithms has proven useful for achieving more accurate and plausible results, as summarised in (Nosrati & Hamarneh, 2016). The main strategies proposed for this problem include incorporating a pre-computed shape prior in the loss function (Zotti *et al.*, 2018), modeling boundaries and edge polarity (Hao *et al.*, 2017), learning the distribution of valid segmentation masks using an autoencoder (Oktay *et al.*, 2017; Painchaud *et al.*, 2020; Gao *et al.*, 2021), and registering an anatomical atlas (Duan *et al.*, 2019; Dong *et al.*, 2020). While such strategies can improve the overall segmentation, they need a sufficient amount of labeled data to learn the shape prior/atlas, which may not be available in practice, or require the custom design of differentiable losses for constraints.

So far, the use of anatomical constraints in semi-supervised segmentation settings remains largely unexplored. To bridge this gap, we propose a Constrained Adversarial Training (CAT) approach for semi-supervised segmentation, that considers complex constraints during training to learn an anatomically-plausible segmentation. While constraints have been exploited in deep learning models for semi-supervised and weakly-supervised segmentation (Kervadec *et al.*, 2019; Peng *et al.*, 2020b; Jia *et al.*, 2017; Pathak *et al.*, 2015), due to the added complexity of optimization, these constraints are typically very simple: region size, adjacency, centroid position, etc. Such constraints are unable to capture complex shape priors of medical images, including region convexity and connectivity, which are highly-relevant for various segmentation tasks. Moreover, incorporating them in a deep learning model requires designing a specific loss function that should be differentiable. In contrast, our CAT method can be used out-of-the-box to add any (non-differentiable) constraint on top of a given segmentation model.

The contributions of our work are as follows:

- We propose a novel framework that obtains anatomically-plausible segmentations by incorporating complex constraints, such as connectivity and convexity, during training. Our framework implements two innovative strategies to exploit such constraints efficiently. First, to include non-differentiable constraints in back-propagation, without having to design a custom loss function, it adopts an optimization approach based on the REINFORCE algorithm that estimates gradients stochastically instead of analytically. Second, to have gradients useful for learning the constraints, it employs an adversarial training strategy that generates on-the-fly examples causing the network to violate the given constraints. By minimizing its error for these adversarial examples, the network can learn how to satisfy the constraints without the need for additional labeled examples.
- To our knowledge, our segmentation method is the first to consider complex anatomical priors in a general semi-supervised setting. In comparison, existing approaches require a large number of labeled images to learn a shape prior (Oktay *et al.*, 2017; Painchaud *et al.*, 2020; Gao *et al.*, 2021) or a complex and problem-specific step involving atlas registration (Duan *et al.*, 2019; Dong *et al.*, 2020). Unlike these approaches, our method needs very few labeled

examples and can be added on top of any segmentation network. The code is available at https://github.com/WangPing521/constraint_aware_vat_semi_supervised_segmentation

3.2 Related work

3.2.1 Semi-supervised Segmentation

Due to the high complexity and cost of generating ground-truth annotations for segmentation, a wide range of semi-supervised methods have been proposed for this problem. Such methods, which focus on exploiting unlabeled data to regularize a model trained with few annotated examples, include approaches based on self-training (Lee *et al.*, 2013), data augmentation (Chaitanya *et al.*, 2019; Zhao *et al.*, 2019), entropy minimization (Vu *et al.*, 2019), adversarial learning (Souly *et al.*, 2017; Zhang *et al.*, 2017, 2020), co-training (Qiao *et al.*, 2018; Peng *et al.*, 2020a; Zhou *et al.*, 2019b), and consistency regularization (Dou *et al.*, 2020; Perone & Cohen-Adad, 2018). The pseudo-label method proposed by Lee et al. (Lee *et al.*, 2013) implements a self-training strategy where a model is first trained with the labeled data and then used to predict labels for unlabelled data. In contrast, Chaitanya et al. (Chaitanya *et al.*, 2019) used data augmentation to generate new training examples for semi-supervised learning. In this approach, a generative model is trained with task-specific data to generate realistic images and corresponding segmentation masks. Entropy minimization methods (Vu *et al.*, 2019) increase the network's confidence for unlabeled examples (i.e., reduce their entropy) to steer the decision boundaries toward low-density regions in prediction space. Virtual Adversarial Training (VAT) (Miyato *et al.*, 2019) generates adversarial perturbations by maximizing the divergence between predictions for original training samples and the corresponding perturbed samples. Training the model with these adversarial examples promotes local distribution smoothness (LDS) which helps the model to become more resilient to noise. Co-training approaches (Blum & Mitchell, 1998) have also shown promising results for semi-supervised segmentation. Peng et al. (Peng *et al.*, 2020a) introduced a deep co-training based method which combines a Jensen–Shannon divergence (JSD) (Englesson & Azizpour, 2021) consistency loss and a model diversity loss using adver-

sarial training. Based on a similar principle, Zhou et al. (Zhou *et al.*, 2019b) trained multiple segmentation networks with different planes of a 3D MRI scan as input, and uses the agreement of these models on unlabeled examples as unsupervised objective. A prominent strategy in recent semi-supervised segmentation approaches is consistency-based regularization (Bortsova *et al.*, 2019). In its simplest form, this strategy imposes a segmentation network to output similar predictions for unlabeled images under different transformations. It is also at the core of temporal ensembling approaches based on Mean Teacher (Perone & Cohen-Adad, 2018; Cui *et al.*, 2019) which enforce the output of a student network at separate training iterations to be similar to that of a teacher network whose parameters are the exponential weighted average of the student's.

In semi-supervised settings where very few labeled images are available, it is usually impossible for a segmentation network to learn the distribution of valid shapes only from labeled images. While regularization approaches can boost performance in these settings, they may not prevent a network from generating anatomically-invalid segmentations. Our CAT method tackles this problem by enforcing the network to satisfy complex anatomical constraints for adversarial perturbations on unlabeled examples.

3.2.2 Constraint-based segmentation

Several works have focused on incorporating constraints in semi-supervised or weakly-supervised segmentation methods (Kervadec *et al.*, 2019; Pathak *et al.*, 2015; Jia *et al.*, 2017; Zhou *et al.*, 2019a). The approach in (Jia *et al.*, 2017) uses a simple L_2 penalty to impose size constraints on segmented regions in histopathology images. Kervadec et al. (Kervadec *et al.*, 2019) leveraged a similar differentiable loss function to impose inequality constraints on the size of segmented regions. Likewise, Zhou et al. (Zhou *et al.*, 2019a) proposed a primal-dual gradient approach minimizing the KL divergence between the class marginal distribution of network predictions and a given target distribution. While demonstrating the benefits of adding constraints in a segmentation model, these methods have two important drawbacks. First, they are limited to simple constraints like region size or centroid position, which are insufficient to characterize

the complex shapes found in medical imaging applications. Second, they require to design a problem-specific differentiable loss and thus have low generalizability. In contrast, by leveraging a stochastic optimization approach based on the REINFORCE algorithm, our method can handle any non-differentiable constraint and be added to any segmentation network without specific requirements.

Recent efforts have also been invested toward adding strong anatomical priors in segmentation networks. In (Oktay *et al.*, 2017), Oktay et al. presented an anatomically-constrained neural network (ACNN) which uses the reconstruction error of an autoencoder on the predicted segmentation as a shape prior. However, since training the autoencoder on segmentation predictions requires a sufficient amount of ground truth masks, this approach is poorly suited to semi-supervised learning settings. The cardiac segmentation approach by Zotti et al. (Zotti *et al.*, 2018) improves accuracy by aligning a probabilistic shape prior computed offline to the predicted segmentation during training. Likewise, Duan et al. (Duan *et al.*, 2019) employed a multi-task approach to locate landmarks which guide an atlas-based label propagation during a refinement step. Despite their added robustness, these approaches also need large annotated datasets to learn the atlas and are sensitive to atlas registration errors. Recently, Painchaud et al. (Painchaud *et al.*, 2020) proposed a segmentation method using a variational autoencoder to learn the manifold of valid segmentations. During inference, predicted segmentations are mapped to their nearest valid point in the manifold. While it offers strong anatomical guarantees, this post-processing method requires pre-computing an important number of valid points. Moreover, the projection of a predicted output on these points can lead to a segmentation considerably different from the ground-truth.

3.3 The proposed method

We start by defining the semi-supervised segmentation setting considered in this work. In this setting, we have a small set of labeled examples $\mathcal{S} = \{(\mathbf{x}_s, \mathbf{y}_s)\}_{s=1}^{|\mathcal{S}|}$ where each $\mathbf{x}_s \in \mathbb{R}^{|\Omega|}$ is an image and $\mathbf{y}_s \in \{0,1\}^{|\Omega|\times|C|}$ is the corresponding ground-truth segmentation mask. Here, $\Omega \subset \mathbb{Z}^2$ denotes the set of image pixels and C is the set of segmentation classes. Moreover,

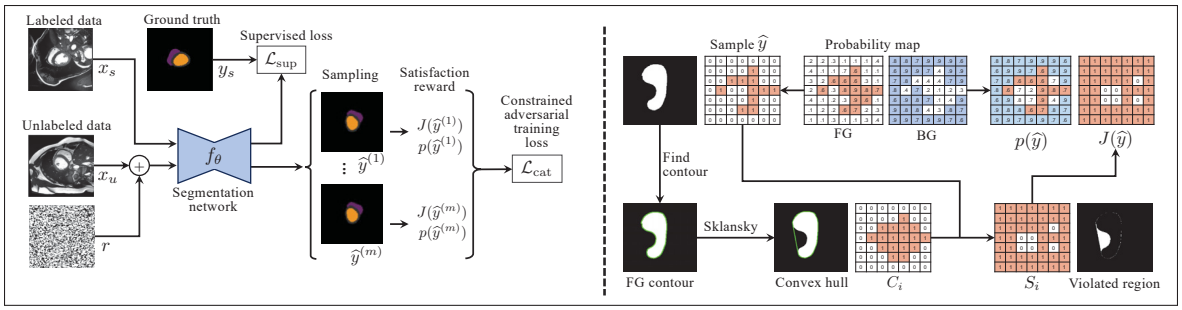


Figure 3.1 **(Left)** Overview of the proposed CAT method. In addition to the supervised loss $\mathcal{L}_{\text{sup}}$ on labeled examples x_s, our model includes a constrained adversarial training loss $\mathcal{L}_{\text{cat}}$ on unlabeled examples x_u on which perturbations r maximizing the prediction divergence and constraint violation are added. **(Right)** Computation of the satisfaction reward for the convexity constraint with $k=1$ (see Section 3.3.3 for details).

we have access to a larger set of unlabeled images $\mathcal{U} = \{\mathbf{x}_u\}_{u=1}^{|\mathcal{U}|}$. The goal is using available examples $\mathcal{D} = \mathcal{S} \cup \mathcal{U}$ to train a segmentation network $f_\theta(\cdot)$, parameterized by a set of weights θ, which can accurately segment unlabeled test images.

While it brings considerable challenges, applying segmentation priors or constraints is often necessary in semi-supervised learning scenarios to get anatomically-plausible predictions. However, many segmentation constraints like region connectivity and convexity are non-differentiable by nature and thus cannot be used in standard optimization algorithms based on back-propagation. To solve this problem, we exploit a stochastic optimization strategy based on the REINFORCE algorithm (Williams, 1992) which computes gradients via sampling. Since few labeled examples are available in our semi-supervised segmentation setting, we exploit unlabeled images with an adversarial training strategy that has two separate goals. The first goal is regularizing the learning, as in VAT (Miyato *et al.*, 2019), by making the network's prediction robust to adversarial perturbations on unlabeled images. The second goal is obtaining useful gradients for learning the constraints. Since non-zero gradients are only obtained when constraints are violated, and constructing examples that lead to invalid predictions is non-trivial for complex constraints, we use adversarial training to generate adversarial examples maximizing the constraint violation loss.

An overview of our CAT method is given in Figure 3.1. A total loss comprised of a supervised loss $\mathcal{L}_{\text{sup}}$ and a constrained adversarial training loss $\mathcal{L}_{\text{cat}}$ is developed:

$$\mathcal{L}_{\text{total}}(\theta; \mathcal{D}) = \mathcal{L}_{\text{sup}}(\theta; \mathcal{S}) + \lambda \mathcal{L}_{\text{cat}}(\theta; \mathcal{U}) \tag{3.1}$$

The supervised loss $\mathcal{L}_{\text{sup}}(\cdot)$ encourages the model to predict segmentation outputs for labeled data that are close to the ground truth. In this work, we use the well-know cross-entropy loss,

$$\mathcal{L}_{\text{sup}}(\theta; \mathcal{S}) = -\frac{1}{|\mathcal{S}|} \sum_{(\mathbf{x},\mathbf{y}) \in \mathcal{S}} \sum_{i \in \Omega} \sum_{j \in C} y_{ij} \log \left[f_\theta(\mathbf{x}) \right]_{ij}. \tag{3.2}$$

The next subsection presents our constrained adversarial training loss that enables the network to learn how to produce valid segmentations for the given constraints.

3.3.1 Constrained adversarial training

Our proposed CAT strategy extends the VAT method (Miyato *et al.*, 2019) enforcing local distribution smoothness (LDS). The VAT regularization loss can be formulated as

$$\mathcal{L}_{\text{vat}}(\theta; \mathcal{U}) = \frac{1}{|\mathcal{U}|} \sum_{\mathbf{x} \in \mathcal{U}} \max_{r,\, \|r\|_2 \leq \epsilon} D_{\text{KL}}\big(f_\theta(\mathbf{x}) \,||\, f_\theta(\mathbf{x}+r) \big) \tag{3.3}$$

where D_{KL} is the KL divergence imposing the prediction for an unlabeled example $\mathbf{x} \in \mathcal{U}$ and its perturbed version $\mathbf{x}+r$ to be similar. Adversarial training seeks to find the perturbation r, whose L_2 norm is bounded by ϵ, maximizing divergence. Thus, minimizing $\mathcal{L}_{\text{vat}}$ makes the distribution locally smooth around unlabeled examples.

While VAT offers an efficient way to regularize learning with unlabeled examples, it does not take into account segmentation constraints, which may lead to predictions that are anatomically invalid. The proposed CAT method addresses this problem by incorporating a constraint

satisfaction loss ℓ_{cons}, measured on segmentation predictions, in the adversarial training process:

$$\mathcal{L}_{\text{cat}}(\theta;\,\mathcal{U}) \;=\; \frac{1}{|\mathcal{U}|}\sum_{\mathbf{x}\in\mathcal{U}}\;\max_{r,\,\|r\|_2\le\epsilon}\Big[D_{\text{KL}}\big(f_\theta(\mathbf{x})\;||\;f_\theta(\mathbf{x}{+}r)\big) \\ +\;\gamma\ell_{\text{cons}}\big(f_\theta(x{+}r)\big)\Big]. \tag{3.4}$$

Hence, adversarial examples try to increase both the prediction divergence and the constraint violation measured by ℓ_{cons}. This has two benefits. First, minimizing $\mathcal{L}_{\text{cat}}$ penalizes invalid predictions explicitly, which further regularizes the learning and guides it toward a valid solution. Second, applying the loss ℓ_{cons} on adversarial examples instead of original images $\mathbf{x}$ increases the chances of violating constraints, which is necessary to obtain useful gradients in the stochastic optimization approach described below.

3.3.2 Stochastic optimization of non-differentiable constraints

Let $\widehat{y}$ be the discrete segmentation predicted by the network for a given image $\mathbf{x}$, i.e., $\widehat{y}_{ij}=1$ if $j=\arg\max_{j'}\,[f_\theta(\mathbf{x})]_{ij'}$ else $\widehat{y}_{ij}=0$. We write the constraint satisfaction loss ℓ_{cons} as a weighted sum of K functions

$$\ell_{\text{cons}}\big(f_\theta(\mathbf{x})\big) \;=\; -\sum_{k=1}^{K}\omega_k\,J_k(\widehat{y}) \tag{3.5}$$

where J_k measures the satisfaction of the k-th constraint and $\omega_k \ge 0$ is a weight controlling the relative importance of this constraint in the loss. To simplify the analysis, without loss of generality, we suppose there is a single constraint satisfaction term J such that $\ell_{\text{cons}}(f_\theta(\mathbf{x})) = J(\widehat{y})$.

We consider two approaches for modeling the constraint satisfaction, *Hard* or *Soft*. For Hard satisfaction, J has a binary output where $J(\widehat{y}) = 1$ if $\widehat{y}$ satisfies the constraint, otherwise $J(\widehat{y}) = 0$. In the case of Soft satisfaction, J outputs a value in the $[0, 1]$ range with 1 corresponding to full satisfaction and 0 to zero satisfaction of the constraint. The latter approach, which also provides a learning gradient for partly-valid solutions, is advantageous when full constraint satisfaction is hard to achieve. Even for soft satisfaction, the constraint function J may not be differentiable. For

instance, determining if a region is convex requires running a non-trivial algorithm which cannot be directly expressed as a function encoded by neural networks. Moreover, as mentioned before, these constraints take a discrete segmentation as input, whereas the output of the segmentation network is a continuous probability. To solve these two problems, we employ a stochastic optimization strategy based on the REINFORCE algorithm (Williams, 1992) used for reinforcement learning. Toward this goal, we consider $\widehat{y}$ as a discrete action sampled over probability $f_\theta(\mathbf{x})$ and J as a reward function. We then maximize the expected reward defined as

$$Q\big(f_\theta(\mathbf{x})\big) \;=\; \mathbb{E}_{\widehat{y}\sim f_\theta(\mathbf{x})}\big[J(\widehat{y})\big]. \tag{3.6}$$

Let $p(\widehat{y}) = \prod_{i\in\Omega} p(\widehat{y}_i)$ be the probability of sampling segmentation $\widehat{y}$ from distribution $f_\theta(\mathbf{x})$, where

$$p(\widehat{y}_i) \;=\; \sum_{j\in C} \widehat{y}_{ij} \cdot [f_\theta(\mathbf{x})]_{ij}, \tag{3.7}$$

the gradient can be estimated as follows:

$$\begin{aligned}
\nabla_\theta Q\big(f_\theta(\mathbf{x})\big) \;&=\; \nabla_\theta\Big[\sum_{\widehat{y}} p(\widehat{y})\, J(\widehat{y})\Big] \\
&=\; \sum_{\widehat{y}} p(\widehat{y})\, J(\widehat{y})\, \nabla_\theta \log p(\widehat{y}) \\
&\approx\; \frac{1}{m}\sum_{s=1}^{m} J(\widehat{y}^{(s)}) \sum_{i\in\Omega} \nabla_\theta \log p\big(\widehat{y}_i^{(s)}\big).
\end{aligned} \tag{3.8}$$

The last line approximates the sum over $\widehat{y}$ with m samples $\widehat{y}^{(s)} \sim f_\theta(\mathbf{x})$. To reduce variance, we follow common practice and center the rewards on their mean $\overline{J}$:

$$\nabla_\theta Q\big(f_\theta(\mathbf{x})\big) \;\approx\; \frac{1}{m}\sum_{s=1}^{m} \big(J(\widehat{y}^{(s)}) - \overline{J}\big) \sum_{i\in\Omega} \nabla_\theta \log p\big(\widehat{y}_i^{(s)}\big). \tag{3.9}$$

Based on this estimation of the gradient, the constraint satisfaction loss for an unlabeled example $\mathbf{x}$ can thus be expressed as

$$\ell_{\text{cons}}\big(f_\theta(\mathbf{x})\big) \;=\; -\frac{1}{m}\sum_{s=1}^{m}\big(J(\widehat{y}^{(s)}) - \overline{J}\big)\sum_{i\in\Omega}\log p\big(\widehat{y}_i^{(s)}\big). \tag{3.10}$$

For some constraints, satisfaction may be difficult to achieve globally. For instance, a constraint imposing a given region to be connected may be violated by a single noisy pixel. Since the discrete segmentations $\widehat{y}$ are sampled randomly from $f_\theta(\mathbf{x})$, they are very unlikely to satisfy this constraint, which leads to a null gradient. To alleviate this problem, we can seek to satisfy the constraint locally, that is, inside smaller regions around each pixel of the image. Instead of a global satisfaction reward $J(\widehat{y})$, we have a local reward $J_i(\widehat{y})$ for each pixel $i \in \Omega$. The constraint loss then becomes

$$\ell^{\text{local}}_{\text{cons}}\big(f_\theta(\mathbf{x})\big) \;=\; -\frac{1}{m}\sum_{s=1}^{m}\sum_{i\in\Omega}\big(J_i(\widehat{y}^{(s)}) - \overline{J}\big)\log p\big(\widehat{y}_i^{(s)}\big). \tag{3.11}$$

The following subsection explains how to use this approach to model two different segmentation constraints.

3.3.3 Examples of non-differentiable constraints

Various anatomical constraints on shape, topology, geometrical or region interaction including containment, exclusion and relative position, and adjacency (Nosrati & Hamarneh, 2016), can be used to improve the segmentation of medical images in semi-supervised settings. Many of these constraints are hard to model as differentiable loss functions for optimization algorithms based on gradient descent. In this paper, we consider three well-known constraints with broad applicability: connectivity, convexity, and horizontal symmetry. While these constraints can be applied to all segmentation classes, for simplicity, we suppose there is a single foreground region to segment.

Connectivity Given a segmented region G, we say that G is connected if and only if there exists a path between each pair of pixels $p, q \in G$ such that all pixels in the path belong to G. In short, this constraint imposes G to have a single connected component. As mentioned above, this constraint is difficult to satisfy globally, especially in early training stages where the network output is noisy. To avoid sparse gradients, we relax the global constraint and instead consider it locally for patches around each pixel $i \in \Omega$. Since satisfaction at each local patch is a necessary condition for global satisfaction, this helps satisfy the constraint over the whole image.

Algorithm 3.1 summarizes the process for computing local connectivity satisfaction $J_i(\widehat{y})$. In a first step, we estimate the *foreground centerness* S_i of each pixel i as the number of foreground pixels in a $\ell \times \ell$ window centered on i. Pixels with a greater number of neighbors are more likely to lie in large components of the foreground. Note that centerness values can be computed efficiently by a simple convolution of the segmentation mask $\widehat{y}$ with a $\ell \times \ell$ kernel of ones, $1_{\ell \times \ell}$. In the second step, we randomly select a pixel i^{seed} with maximum centerness and run the flood-fill algorithm using this pixel as starting seed. This finds all pixels that are connected to i^{seed} by a path in the foreground, in time linear to the number of pixels. The last step evaluates the local connectivity at each pixel i considering a $k \times k$ window centered on i. Let $C \in \{0, 1\}^{|\Omega|}$ be a binary map measuring the connectivity of each pixel to i^{seed}. A pixel i *violates* connectivity if it is part of the foreground but not connected to i^{seed}. The satisfaction of i can thus be expressed as

$$S_i \;=\; \mathbb{1}(\widehat{y}_i = 0 \;\vee\; C_i = 1), \tag{3.12}$$

where $\mathbb{1}(\cdot)$ is the indicator function. To measure local connectivity every pixel, we count the number of satisfying pixels N in the surrounding window, which can be obtained efficiently by convolving S with a $k \times k$ kernel of ones, $1_{k \times k}$. Hard constraint is achieved if and only if all pixels in the window are satisfying: $J_i(\widehat{y}) = \mathbb{1}\big(N_i = k^2\big)$. On the other hand, soft satisfaction measures the ratio of satisfying pixels: $J_i(\widehat{y}) = N_i / k^2$.

Convexity A segmented region G is said to be convex if, for any two pixels $p, q \in G$, the line from p to q falls entirely in G. For this constraint, we first find the contour pixels of the

Algorithm 3.1 Computation of the local connectivity satisfaction reward

Input: The segmentation $\widehat{y}^{(s)}$ sampling from output probability distribution of the model f_θ, $s = 1, \ldots, m$
Output: Reward $J(\widehat{y}^{(s)})$, probability of sampling segmentation $p\big(\widehat{y}_i^{(s)}\big)$

Compute the foreground centerness of each foreground pixel via convolution: $F^{(s)} = \widehat{y}^{(s)} * 1_{\ell \times \ell}$;
Randomly select i^{seed} in $\arg\max_i F_i^{(s)}$;
Run the flood-fill algorithm starting at pixel i^{seed} to obtain the binary map of connected pixels $C^{(s)}$;
Compute pixel-wise satisfaction map:
  $S = \mathbb{1}\big(\widehat{y} = 0 \ \vee \ C = 1\big)$;
Compute the local number of satisfying pixels:
  $N = S * 1_{k \times k}$;
if *Constraint is* Hard **then**
  $J(\widehat{y}^{(s)}) = \mathbb{1}\big(N = k^2\big)$;
end if
if *Constraint is* Soft **then**
  $J(\widehat{y}^{(s)}) = N/k^2$;
end if
Compute $p\big(\widehat{y}^{(s)}\big)$ using Eq. (3.7);
return $J(\widehat{y}^{(s)})$, $p\big(\widehat{y}^{(s)}\big)$;

foreground G and then compute their convex hull using Sklansky's algorithm (Graham & Yao, 1983). The complexity of this step is in $O(n \log n)$ where n is the number of contour points of G. An approach similar to the connectivity constraint is employed to compute the rewards $J(\widehat{y})$. In this case, the binary map C is such that $C_i = 1$ if pixel i is inside the convex hull of the foreground, otherwise $C_i = 0$. Then, pixel i *violates* the convexity constraint if it is in the convex hull but was labeled as background. The satisfaction for i can thus be defined as

$$S_i \ = \ \mathbb{1}\big(\widehat{y}_i = 1 \ \vee \ C_i = 0\big). \tag{3.13}$$

The hard and soft satisfaction reward for each pixel i is then computed as before, based on the number of satisfying pixels in a $k \times k$ window centered on i. The process of computing the reward and probability map, $J(\widehat{y})$ and $p(\widehat{y})$, is illustrated in the right of Figure 3.1.

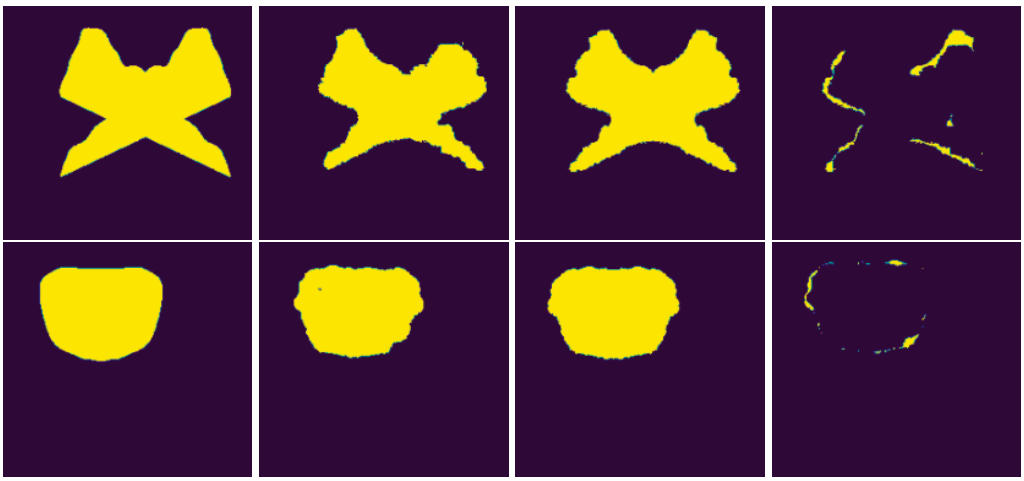


Figure 3.2 Visualization of the prediction, corresponding symmetric shape, and symmetry violation map.

Horizontal symmetry Given a segmented region G, we say that G is horizontally symmetric if it can be divided in two identical halves by a vertical line. The horizontal symmetry of a segmentation with respect to a vertical line is assessed by verifying that each pixel predicted as foreground on either side of this line has a corresponding foreground pixel on the other side of the line. Let C be the symmetric binary map obtained by projecting foreground pixels across the symmetry line (see Figure 3.2). A pixel i *violates* the symmetry constraint if it is inside the symmetric map but was labeled as background. The satisfaction of i is thus given by

$$S_i \;=\; \mathbb{1}\big(\widehat{y}_i = 1 \;\vee\; C_i = 0\big). \tag{3.14}$$

The complexity of this constraint comes from the fact that the true line of symmetry is unknown, and using a wrong line will result in a symmetric but incorrect shape. To solve this problem, we compute the foreground's horizontal center of mass c_x, and then consider multiple lines around this center, with x-coordinate $c_x+5\delta$, $\delta=\{-3,-2,-1,0,1,2,3\}$. For computing the satisfaction reward, we keep the line which violates the least number of pixels.

3.3.4 Reverse reward formulation

Our stochastic optimization approach of Eq. (3.10) is based on a reinforcement learning algorithm where actions (i.e., segmentation) sampled from a distribution receive a reward (i.e., satisfaction of constraints). A problem which may arise from this formulation is that, when the segmentation is confident (i.e., values of $f_\theta(\mathbf{x})$ are near to 0 or 1), the gradient will be small since $\log p(\widehat{y}_i) \approx 0$ for most pixels i. A common solution for this problem in reinforcement learning is to add an entropy regularization term that prevents the policy from becoming too "deterministic". However, even when increasing entropy, a large number of samples may be needed to find high-reward segmentations having a low probability. We address this issue with an alternative formulation named *Reverse reward* which, instead of rewarding selected actions, penalizes non-selected ones. The modified loss for this formulation is as follows:

$$\ell_{\text{cons}}^{\text{rev}}\big(f_\theta(\mathbf{x})\big) \;=\; \frac{1}{m}\sum_{s=1}^{m}\sum_{i\in\Omega}\big(J_i(\widehat{y}^{(s)}) - \overline{J}\big)\log\big(1 - p(\widehat{y}_i^{(s)})\big). \tag{3.15}$$

Note that this strategy is similar to the Non-saturating Minimax loss employed to trained GANs (Goodfellow *et al.*, 2016), where the generator loss $\log\big(1-D(G(z))\big)$ is replaced by $-\log D(G(z))$.

3.4 Experiments

We first present the datasets and metrics used to evaluate the proposed method. We then provide information on the experimental setup, including implementation details and compared approaches.

3.4.1 Datasets and metrics

Datasets

In a first step, synthetic data is used to demonstrate our method's ability to include connectivity, convexity, and horizontal symmetry constraints in segmentation. To further assess our method's

performance, experiments are then conducted on four clinically-relevant benchmark datasets: the Automated Cardiac Diagnosis Challenge (ACDC) (Bernard *et al.*, 2018), the 2012 Prostate MR Image Segmentation (PROMISE12) Challenge dataset (Litjens *et al.*, 2014), Prostate sub-task dataset of the Medical Segmentation Decathlon Challenge (Antonelli *et al.*, 2022), and Hippocampus sub-task dataset of the Medical Segmentation Decathlon Challenge (Antonelli *et al.*, 2022).

Synthetic data: We generated three synthetic datasets, a first one for the segmentation scenario involving connectivity constraints, another for the scenario with convexity constraints, and the last one for horizontal symmetry constraints. Each dataset contains 50 images of size 256×256, which are split into a training set of 40 images and a test set of 10 images. Images were created by superimposing a foreground region on a background with a different mean intensity, and then adding Gaussian noise on the resulting image. Superimposed foregrounds may satisfy or not the constraints. Examples of synthetic images are shown in Figures 3.3, 3.4, and 3.5.

ACDC: This dataset consists of 200 MRI scans from 100 patients, including 20 healthy patients, 20 patients with previous myocardial infarction, 20 patients with dilated cardiomyopathy, 20 patients with an hypertrophic cardiomyopathy, and 20 patients with abnormal right ventricle. Scans correspond to end-diastolic (ED) and end-systolic (ES) phases, and were acquired on 1.5T and 3T systems with resolutions ranging from 0.70×0.70 mm to 1.92×1.92 mm in-plane and 5 mm to 10 mm through-plane. Three cardiac regions are labeled in the ground-truth: left ventricle (LV), right ventricle (RV) and myocardium (Myo). In our experiments, we used a split of 75 subjects (150 scans) for training, 5 subjects (10 scans) for validation, and 20 subjects (40 scans) for testing. We slice 3D-MRI scans into 2D images which are then randomly cropped to a size of 192×192.

PROMISE12: This dataset comprises multi-centric transversal T2-weighted MR images of prostates from 50 subjects. Volumetric images were acquired with multiple MRI vendors and different scanning protocols, and are thus representative of typical MR images acquired in a clinical setting. Image resolution ranges from 15×256×256 to 54×512 ×512 voxels with a

spacing ranging from 2×0.27×0.27 to 4×0.75× 0.75 mm^3. A single prostate region is labeled in the ground-truth. We slice volumetric images to 2D images along short-axis and then randomly crop images into input patches of size 192×192. The data was split as follows: 40 subjects for training set and 10 subjects for testing.

Medical Segmentation Decathlon Prostate dataset: This third dataset consists of 48 multi-parametric MRI (32 MRIs are labeled) studies provided by Radboud University. Two prostate structures are labeled in the ground-truth: peripheral zone (PZ) and central gland (CG). We used a split of 24 examples for training and 8 for testing. A 3-fold cross validation on the training set is conducted to determine the hyper-parameters.

Medical Segmentation Decathlon Hippocampus dataset: This dataset consists of 394 T1-weighted MRI (263 MRIs are labeled) acquired at the Vanderbilt University Medical Center. Two structures are labeled in the ground-truth: anterior and posterior of hippocampus. We used a split of 224 examples for training, 13 for validation, and 26 for testing. We slice 3D-MRI scans into 2D images which are then resized to a size of 64×64.

Metrics

Three general performance metrics, dice similarity coefficient (DSC), modified Hausdorff distance (MHD) and average symmetric surface distance (ASSD), are employed to evaluate the proposed methods. To compute these metrics, we reconstruct the segmented volume by concatenating masks predicted for individual slices. Additionally, non-connectivity (N-conn) and non-convexity (N-conv) measures are used to evaluate the constraint satisfaction of the proposed method.

Dice: The DSC measures the degree of overlap between the segmentation region and ground truth. It is defined as

$$\mathrm{DSC}(G, S) = \frac{2|S \cap G|}{|S| + |G|}, \tag{3.16}$$

where S is the predicted labels and G is the corresponding ground truth labels. DSC values range from 0 to 1, a higher value representing a better segmentation.

MHD: This boundary distance metric is derived from Hausdorff distance (HD) that measures the largest distance between a point in segmentation and its nearest point in the ground truth. The MHD, which is more robust to outliers (Dubuisson & Jain, 1994), is defined as

$$\mathrm{MHD}(G,S) \;=\; \max\left\{d(G,S),\, d(S,G)\right\} \tag{3.17}$$

where $d(S,G)=\frac{1}{N_s}\sum_{s\in S} d(s,G)$. A smaller MHD value indicates a better segmentation.

ASSD: This metric is defined as the average of nearest distances from points on the segmented region boundary to points on the ground truth boundary, and vice versa:

$$\mathrm{ASSD}(G,S) \;=\; \frac{1}{N_s+N_g}\Big(\sum_{g\in G} D_S(g) + \sum_{s\in S} D_G(s)\Big) \tag{3.18}$$

where $D_S(g) = \min_{s\in S} \|g-s\|$.

N-conn: This measures the satisfaction of the connectivity constraint as the proportion of pixels predicted as foreground which are not connected to the pixel with highest centerness (see Section 3.3.3). An N-conn of 0 is obtained when the predicted segmentation is fully-connected.

N-conv: Similarly, convexity satisfaction is measured as the proportion of pixels predicted as background which are in the convex hull of the foreground region. An N-conv of 0 means the predicted foreground is convex.

3.4.2 Experimental setup

Network and hyper-parameters We used the popular light-weight architecture Enet (Adam *et al.*, 2016) as the underlying segmentation network, as it offers a good trade-off between accuracy and amount of parameters. This architecture employs a convolutional block with short skip connections, called bottleneck block, and is comprised of different stages: an initial stage of regular convolutions, followed by stages with different numbers of bottleneck blocks, and a final stage of 1×1 convolutions to generate the final segmentation probability map. All experiments

were carried out using a rectified Adam optimizer and a learning rate warm-up strategy based on cosine decay (Loshchilov & Hutter, 2016). For all datasets, the learning rate was initialized to 1×10^{-4}, increased by a factor of 300 in the first 10 epochs and then decreased with a cosine scheduler for the following 90 epochs.

A grid search strategy measuring performance on the validation set was employed for selecting the main hyperparameters of our method. Table 3.1 gives the values of hyperparameter λ controlling the weight of our constrained adversarial training loss ℓ_{cat} in the overall loss of Eq. (3.1), hyperparameter γ controlling of the trade-off between local distribution smoothness (LDS) and constraint satisfaction in ℓ_{cat} in Eq. (3.4), and hyperparameter ϵ controlling the level of adversarial perturbation in Eq. (3.4), for the ACDC, Promise12, Prostate, and Hippocampus datasets. For connectivity constraints, a 5×5 window was used to measure foreground centerness (i.e., $\ell = 5$) and a 4 neighbor connectivity employed in the flood-fill algorithm (see Algorithm 3.1). For both connectivity and convexity constraints, we used $m = 10$ discrete samples in the stochastic optimization and a 3×3 window to measure local satisfaction (i.e., $k = 3$).

Compared approaches We compared our proposed method with several state-of-art approaches and baselines for semi-supervised segmentation. Since our plug-in method can be added on top of any segmentation algorithm, we also evaluated its combination with two popular semi-supervised learning approaches, Mean Teacher and Co-training. A brief description of tested approaches is given below.

Baseline: This approach trains the network using only the supervised loss on labeled images. When considering all training images as labeled, this gives a fully-supervised *upper bound* on performance. Conversely, in a setting where only few training images are labeled, this baseline represents a lower bound on performance since unlabeled images are ignored.

Entropy min (Vu *et al.*, 2019): In addition to the supervised loss on labeled data, this well-known semi-supervised method also minimizes the pixel-wise entropy of predictions for unlabeled images. Making the network's predictions to be more confident regularizes the learning by forcing the decision boundary to pass through low-density regions of the data space.

Table 3.1 Hyper-parameter setting of our CAT method and its variants, for the ACDC, PROMISE12 and Prostate datasets.

	Method	λ	γ	ϵ
ACDC	CAT (no adv)	–	7×10^{-5}	–
	CAT	5×10^{-3}	5×10^{-5}	0.5
	CoT + CAT	5×10^{-3}	5×10^{-5}	0.5
	MT + CAT	2	5×10^{-5}	0.5
PROMISE12	CAT (no adv)	–	1×10^{-3}	–
	CAT	5×10^{-3}	5×10^{-4}	1.0
	CoT + CAT	3×10^{-4}	1×10^{-3}	1.0
	MT + CAT	5×10^{-2}	1×10^{-5}	1.0
Prostate	CAT (no adv)	–	5×10^{-6}	–
	CAT	5×10^{-4}	7×10^{-6}	0.1
	CoT + CAT	5×10^{-3}	5×10^{-5}	0.1
	MT + CAT	2	5×10^{-5}	0.1
Hippocampus	CAT (no adv)	–	1×10^{-5}	–
	CAT	1×10^{-3}	1×10^{-5}	0.1
	CoT + CAT	5×10^{-3}	5×10^{-5}	0.1
	MT + CAT	4	5×10^{-3}	0.1

VAT (Miyato *et al.*, 2019): This semi-supervised approach optimizes a minimax problem where adversarial perturbations are added on training examples to maximize the prediction divergence of the network, while the network parameters are updated to minimize this divergence.

Co-training (Qiao *et al.*, 2018): This approach jointly trains two models that improve their individual performance by exchanging information during training. Besides the supervised loss on labeled data, co-training employs a consistency loss based on JSD encouraging the models to give similar predictions for unlabeled data.

Mean Teacher (Antti & Valpola, 2017): Mean Teacher adopts a teacher-student framework where two networks sharing the same architecture learn from each other. Given an unlabeled image, the Student model seeks to minimize the prediction difference with the Teacher network whose weights are a temporal exponential moving average (EMA) of the student's. Following the standard practice, we fix the decay coefficient in EMA to be 0.999.

AE-prior (Oktay *et al.*, 2017) (Gao *et al.*, 2021): This method trains an autoencoder (AE) to reconstruct the predicted and ground-truth segmentation maps. The latent features of the AE are then used as shape prior while training the segmentation network. This is achieved by minimizing the L2 distance between the latent features of the segmentation prediction for a given image and those of its corresponding ground-truth mask. In our experiments, we followed the implementation of (Gao *et al.*, 2021) which uses an adversarial shape loss where the feature distance is minimized for the segmentation network but maximized for the AE. This adversarial approach encourages the AE to capture subtle differences between real and predicted shapes. As the shape loss requires labeled examples, it cannot be employed by itself in a semi-supervised setting. Hence, to have a fair comparison, we also added an entropy minimization loss computed on unlabeled examples.

CAT: Our proposed method using the objective function of Eq. (3.1). As explained in Section 3.3.1, adversarial examples are generated to maximize both the network's prediction divergence and the constraint satisfaction loss. Minimizing the loss for these examples regularizes the training and helps the network learn to satisfy the constraints.

CAT (no adv): This variant of our method, achieved by setting $\epsilon = 0$ in Eq. (3.4), disables the adversarial training. As a result, the prediction divergence term is ignored and the constraint loss ℓ_{cons} is applied on unlabeled examples without perturbations.

CoT + CAT: In this plug-in variant, we add our CAT method on top of Co-training. In the modified model, the first network works as in the original co-training approach, and our CAT loss is added to the second one. Following Co-training, prediction consistency is enforced with a JSD-based loss.

MT + CAT: This other plug-in variant adds our method to Mean Teacher. Similar to the previous variant, we add the CAT loss to the Student network and minimize the MSE between the student and teacher's predictions. The Teacher parameters are updated following standard Mean Teacher.

To have a fair comparison, all tested methods were implemented in a single framework where the same segmentation backbone and grid search hyperparameter-tuning strategy were used for all compared approaches.

3.5 Results

3.5.1 Experiments on synthetic data

Ablation study We start with an ablation study evaluating the usefulness of our method's main strategies and components. This study considers the scenario with connectivity constraints, using no adversarial perturbation ($\epsilon = 0$). We then compare the performance of the Hard and Soft satisfaction strategies described in Algorithm 3.1, as well as the Standard and Reverse reward formulations of Equations (3.11) and (3.15).

Table 3.2 reports the results of this ablation study. We see that using a Soft satisfaction in training boosts segmentation accuracy and connectivity on the test set, with a 0.34–0.48% higher DSC and 1.86–3.26% more connected pixels (lower N-conn) compared to the Hard satisfaction approach. Likewise, employing the Reverse formulation that penalizes constraint violation instead of rewarding satisfaction leads to a better performance, increasing DSC by 2.71–2.85% and pixel connectivity by 2.52–3.92%. The Soft satisfaction approach with Reverse reward formulation achieves the best overall performance with a DSC of 89.22% and N-conn of 9.96%. Hence, we use this model for remaining experiments.

Next, we assess the impact of the constraint satisfaction loss ℓ_{cons} in Eq. (3.4) by varying its importance weight γ. As can be seen in Table 3.3, increasing λ up to 0.00005 improves accuracy and constraint satisfaction consistently with an increased DSC and reduced N-conn. However, using a too large γ hurts performance, even though a comparable N-conn is obtained. This is because giving exaggerated importance to constraint satisfaction can make the network optimization unstable.

Table 3.2 DSC (%) and N-conn (%) of our method with different ablation settings on connectivity synthetic dataset. We report the mean and stdev. obtained over three runs.

Soft	Reverse	DSC (%)	N-conn (%)
✗	✗	86.03 ± 3.02	15.74 ± 4.62
✗	✓	88.88 ± 1.35	13.22 ± 1.81
✓	✗	86.51 ± 0.47	13.88 ± 0.65
✓	✓	**89.22 ± 1.26**	**9.96 ± 1.03**

Table 3.3 DSC (%) and N-conn (%) of our method when take vary constraint weights on connectivity synthetic dataset. We report the mean and stdev. obtained over three runs.

γ	DSC (%)	N-conn (%)
0.001	85.02 ± 0.76	10.23 ± 2.70
0.0005	84.91 ± 1.70	10.59 ± 0.43
0.0001	87.35 ± 0.28	10.85 ± 2.19
0.00005	**89.22 ± 1.26**	**9.96 ± 1.03**
0.00001	89.12 ± 0.57	12.53 ± 1.55

Visualization of constraint satisfaction Figures 3.3 and 3.4 show examples of the segmentation obtained at different training stages for the scenarios with connectivity and convexity constraints. As can be seen, the Baseline method which does not consider constraints during training fails yields segmentation predictions that are either disconnected or non-convex, despite having a high overlap with the ground-truth. In contrast, our CAT method converges to fully-connected and convex segmentations. Figure 3.5 shows examples of segmentations for the scenario with horizontal symmetry constraints. We see that the proposed method generates better segmentations than the baseline model without symmetry constraints, demonstrating its efficiency.

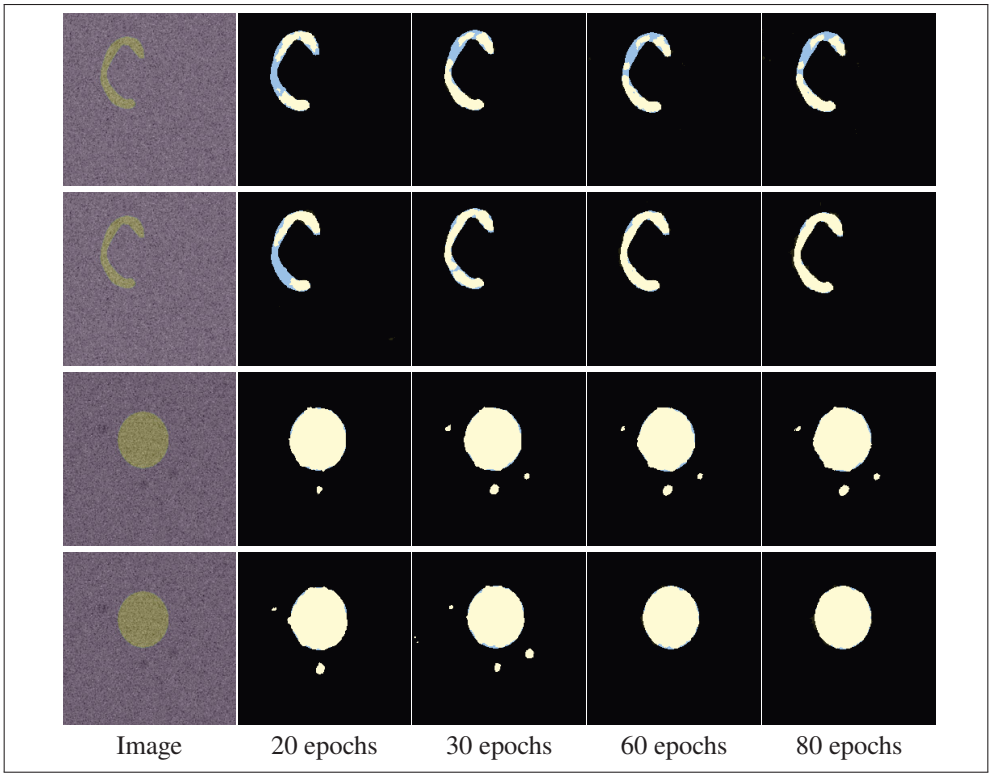


Figure 3.3 Example of segmentation with *connectivity* constraints during training. First and third rows are predictions of the Baseline, second and last rows are those of our CAT method. Blue regions represent the ground truth and overlaid yellow ones are the predicted segmentation.

3.5.2 Experiments on benchmark datasets

Size of local satisfaction kernel The proposed method evaluates the satisfaction of constraints in local regions of size $k \times k$. For the hard satisfaction approach, using a larger $k$imposes a more strict satisfaction, since a single non-satisfied pixel in the region causes the constraint to be violated, but selecting a too large k can result in non-informative gradients if the constraint is violated in all regions. The soft satisfaction strategy avoids this problem by enabling partial constraint satisfaction, however, employing a larger k in this strategy causes local information to be lost (the $k \times k$ kernel acts as a mean filter). To analyze the impact of kernel size k on performance, we tested our soft satisfaction method on the task of cardiac segmentation with connectivity constraints, for different values of k. As reported in Table 3.4, using a smaller

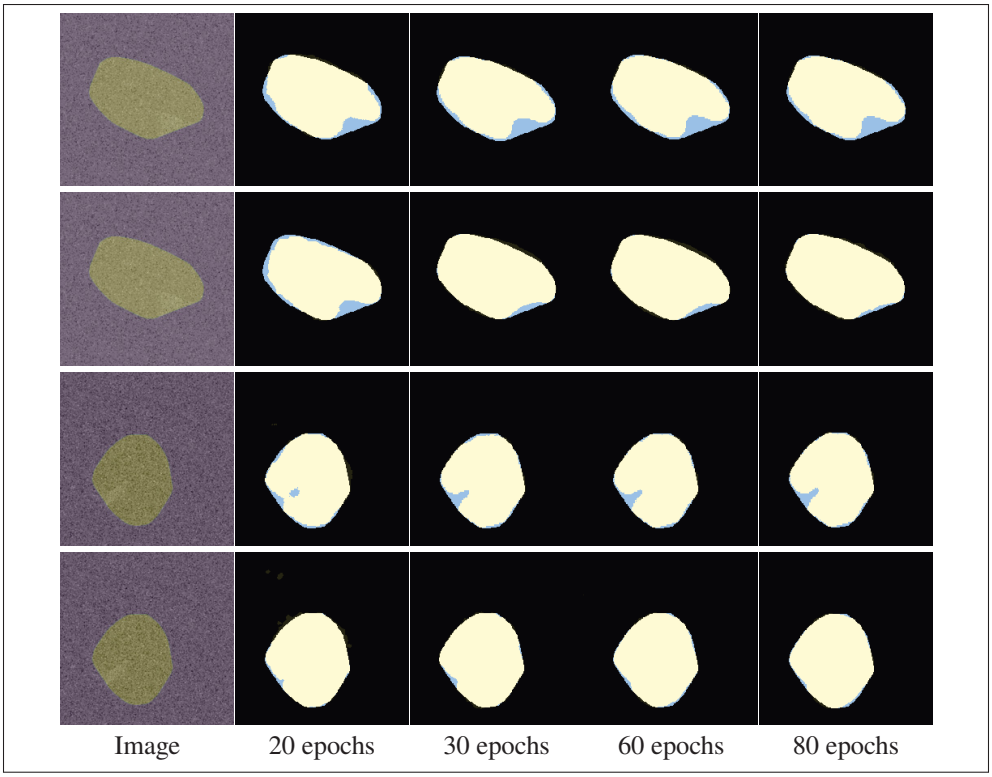


Figure 3.4 Example of segmentation with *convexity* constraints during training. First and third rows are predictions of the Baseline, second and last rows are those of our CAT method. Blue regions represent the ground truth and overlaid yellow ones are the predicted segmentation.

Table 3.4 Impact of local satisfaction kernel size k.

k	**DSC** (%)	**N-conn** (%)
3	**83.77** ± **0.17**	**10.25** ± **0.99**
5	81.43 ± 0.15	10.71 ± 0.49
7	80.49 ± 0.50	10.75 ± 1.37
11	79.81 ± 0.20	11.34 ± 1.13

kernel size leads to a better performance in terms of both segmentation accuracy and constraint satisfaction. This is due to the fact that larger values for k produce reward maps that penalize predictions for thin structures like the myocardium (see Figure 3.6), whether these predictions are correct or not.

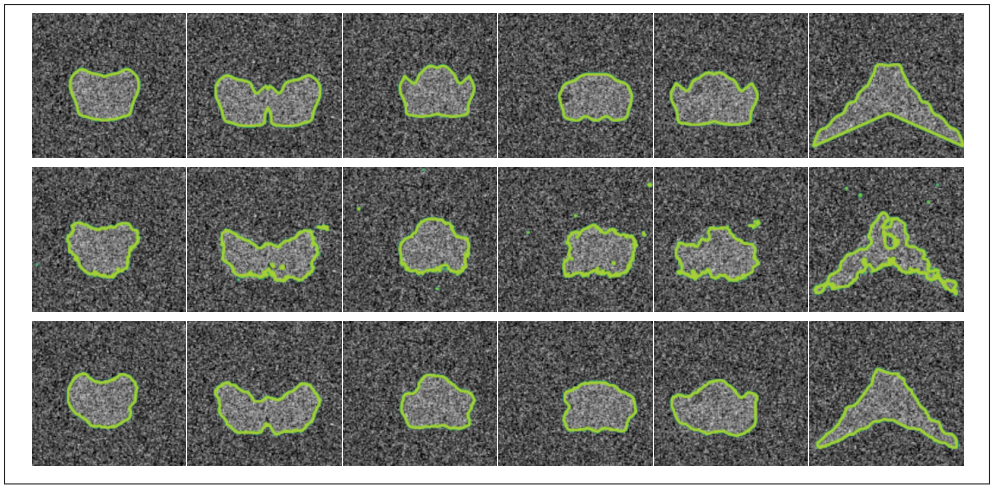


Figure 3.5 Visualization of segmentation with horizontal symmetry. The first row shows the ground-truth, the second row the segmentation of the baseline trained only on labeled data, and the last row the results of our method without adversarial training.

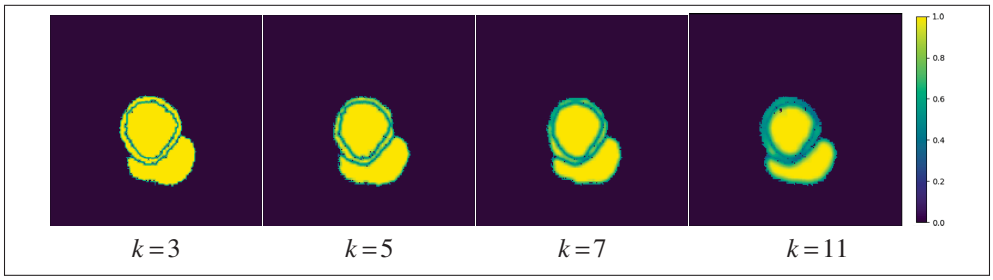


Figure 3.6 Soft reward maps with different kernel sizes k for the local satisfaction.

Comparison between hard and soft constraint satisfaction To show the advantage of soft satisfaction over hard satisfaction, we compare the two strategies for connectivity constraints on the ACDC dataset. Figure 3.7 (left) shows the predictions and corresponding rewards of hard and soft constraints. In contrast to the binary reward of the hard constraint, different reward values are assigned to boundary pixels for the soft constraint strategy, which provides a learning gradient for partly-valid solutions. The benefit of using a soft satisfaction strategy during training can also be appreciated in Figure 3.7 (right) showing the connectivity satisfaction (N-conn) of

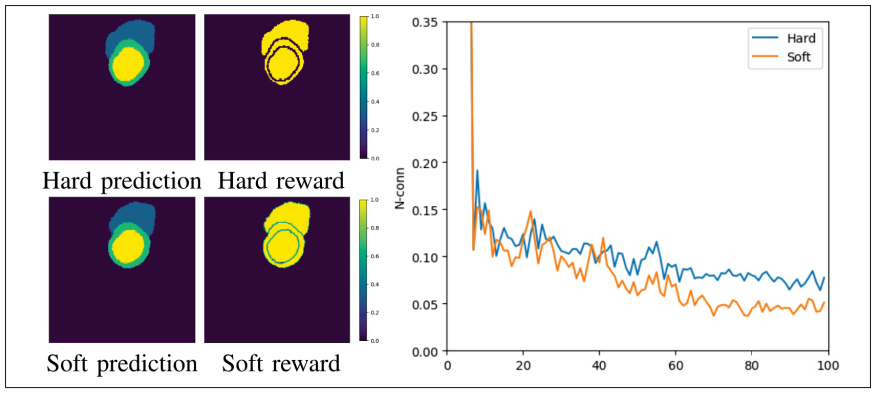


Figure 3.7 The prediction and corresponding reward map (left), and the connectivity satisfaction curve in training stage (right).

test examples at different training epochs. As can be observed, this strategy reduces the N-conn faster than the hard constraint one, and convergences to a lower value.

Analysis of adversarial training loss terms Next, we perform an ablation study to investigate the benefits brought by the two adversarial training loss terms of Eq. (3.4), i.e., the KL divergence loss (D_{KL}) of VAT enforcing local distribution smoothness, and the proposed constraint loss ℓ_{cons}. For the latter, we test three different variants computing ℓ_{cons} on original unlabeled images ($\mathbf{x}_u$), adversarial images ($\mathbf{x}_u + r$) or both. Table 3.5 shows results of this ablation study for connectivity constraints on the ACDC data with a labeled data ratio of 5%. As can be seen, having a KL-based consistency term is essential to generate good adversarial examples. Moreover, applying the constraint loss on these adversarial examples instead of original images improves both segmentation accuracy (DSC) and constraint satisfaction (N-conn). However, using the constraint loss on both original images and adversarial examples gives no additional benefit, while increasing computational costs.

Comparison of VAT and CAT as plug-in As mentioned before, our CAT method can be added as plug-in on top of any semi-supervised segmentation algorithm. We tested two variants of our method combining it to Co-training (CoT + CAT) and Mean Teacher (MT + CAT). For comparison, we also applied VAT on top of Co-training (CoT + VAT) and Mean Teacher (MT + VAT). Table 3.6 shows the results of the semi-supervised baselines and their variants for

Table 3.5 Ablation experiments on the KL divergence and constraint loss terms of Eq. (3.4).

D_{KL}	ℓ_{cons}		**DSC** (%)	**N-conn** (%)
	x_u	x_u+r		
	✓		80.40 ± 0.65	14.62 ± 1.57
✓	✓		83.52 ± 0.44	10.81 ± 0.33
✓		✓	83.77 ± 0.17	10.25 ± 0.99
✓	✓	✓	83.47 ± 0.59	10.70 ± 0.51

connectivity constraints on ACDC, Promise and Prostate, as well as convexity constraints on the LV class of ACDC. For each segmentation task and performance metric, we use bold font values to identify the best method and underline these values if they represent a statistical significant improvement ($p < 0.05$ in a one-tailed paired t-test). As we can see, using our CAT method as plug-in outperforms VAT in terms of both DSC and constraint satisfaction for all tasks and semi-supervised baselines except for DSC in the LV task where Mean Teacher is better.

Comparison to the state-of-art We compared our method against various baselines and state-of-art approaches for semi-supervised segmentation on the ACDC, Promise12, Prostate, and Hippocampus datasets. For each dataset, we train the compared methods with a varying number of labeled examples to simulate different levels of supervision: 100%, 5% and 3% for ACDC, 100%, 8% and 5% for Promise12, 25% for Prostate, and 5% for Hippocampus. Using fewer labeled examples led to poor results for all methods. Results for the segmentation scenario with connectivity constraints are summarized in Tables 3.7 (ACDC), 3.8 (Promise12), 3.9 (Prostate), and 3.10 (Hippocampus). Compared to the Baseline model trained using only the supervised loss ℓ_{sup} on labeled data, our CAT method achieves a significant increase in segmentation accuracy and pixel connectivity. For ACDC, it obtains a 8.85% higher DSC with a 7.36% lower N-conn than the Baseline when considering 3% of training examples as labeled. Similarly, it boosts DSC by 10.76% and connectivity by 17.96% for Promise12 with 5% labeled examples.

Table 3.6 Ablation results generated by variants of VAT and CAT on top of Co-training and Mean Teacher. We report the mean and stdev. obtained over three runs.

Label %	Method	DSC (%)	MHD (mm)	ASSD (mm)	N-conn (%)
3 % ACDC	Co-training	76.22 ± 2.33	1.58 ± 0.20	1.06 ± 0.13	19.74 ± 3.30
	CoT + VAT	76.13 ± 1.89	0.92 ± 0.11	0.70 ± 0.06	18.84 ± 0.68
	CoT + CAT	**76.95 ± 1.35**	**0.90 ± 0.01**	**0.69 ± 0.03**	**18.38 ± 0.82**
	Mean Teacher	82.39 ± 0.48	1.13 ± 0.12	0.77 ± 0.06	8.45 ± 0.49
	MT + VAT	80.91 ± 0.82	**0.85 ± 0.06**	**0.64 ± 0.04**	8.79 ± 0.80
	MT + CAT	**82.41 ± 0.92**	1.11 ± 0.08	0.77 ± 0.05	**7.97 ± 1.85**
5 % PROMISE	Co-training	52.60 ± 0.67	9.78 ± 1.79	6.54 ± 1.05	34.59 ± 2.59
	CoT + VAT	52.42 ± 0.72	**4.16 ± 0.11**	**2.99 ± 0.05**	37.68 ± 3.10
	CoT + CAT	**67.75 ± 0.99**	5.16 ± 1.18	3.50 ± 0.53	**11.23 ± 3.32**
	Mean Teacher	75.95 ± 4.20	3.43 ± 1.62	2.59 ± 1.03	15.30 ± 8.52
	MT + VAT	77.17 ± 1.97	3.38 ± 0.29	2.43 ± 0.16	11.16 ± 6.50
	MT + CAT	**77.42 ± 3.53**	**2.99 ± 0.88**	**2.23 ± 0.50**	**10.47 ± 6.76**
25 % Prostate	Co-training	63.29 ± 0.37	2.52 ± 0.15	2.07 ± 0.16	**18.58 ± 1.34**
	CoT + VAT	63.28 ± 0.40	**2.35 ± 0.10**	2.01 ± 0.13	20.35 ± 0.79
	CoT + CAT	**63.67 ± 0.62**	2.39 ± 0.29	**1.89 ± 0.17**	21.58 ± 1.50
	Mean Teacher	65.99 ± 2.50	2.43 ± 0.73	2.00 ± 0.52	15.82 ± 0.99
	MT + VAT	66.22 ± 2.75	2.83 ± 1.01	2.15 ± 0.69	16.20 ± 3.01
	MT + CAT	**66.76 ± 2.71**	**2.35 ± 0.41**	**1.82 ± 0.27**	**14.91 ± 1.26**
3 % LV	Co-training	83.20 ± 0.17	1.81 ± 0.46	1.19 ± 0.23	4.23 ± 0.47
	CoT + VAT	83.31 ± 1.36	1.84 ± 0.44	1.20 ± 0.22	3.93 ± 0.69
	CoT + CAT	**84.89 ± 0.39**	**0.89 ± 0.05**	**0.72 ± 0.04**	**3.91 ± 0.08**
	Mean Teacher	**88.99 ± 0.92**	**1.29 ± 0.40**	**0.81 ± 0.22**	4.12 ± 0.35
	MT + VAT	88.32 ± 1.61	1.75 ± 0.96	1.04 ± 0.47	3.74 ± 0.20
	MT + CAT	87.21 ± 0.52	1.38 ± 0.19	0.89 ± 0.10	**3.61 ± 0.14**

The usefulness of our adversarial training strategy can also be appreciated by comparing the results of CAT with the CAT (no adv) setting without this strategy. For ACDC with 3% labeled examples, employing adversarial training (CAT) increases DSC by 10.24% while reducing non-connected pixels by 6.71%. Although less pronounced, improvements in accuracy and connectivity are also obtained for the PROMISE12 dataset with 5% of training examples being labeled. In contrast, AE-prior, a shape-constrained method that relies on labeled data, obtains worse performance due to the lack of labeled data for properly training the autoencoder. As for our CAT method, adding it to the algorithms improves both DSC and N-conn in *all but one* case (N-conn of CoT + CAT for Prostate with 25%). Furthermore, the combination of our method

Table 3.7 DSC, MHD, ASSD and non-connectivity (N-conn) for segmenting the ACDC. We report the mean and stdev. obtained over three runs.

Label %	Method	DSC (%)	MHD (mm)	ASSD (mm)	N-conn (%)
100 %	Baseline	89.74 ± 0.42	0.40 ± 0.05	0.33 ± 0.03	8.37 ± 0.29
5 %	Baseline	80.15 ± 0.74	1.18 ± 0.06	0.82 ± 0.06	11.82 ± 0.46
	Entropy min	80.53 ± 1.06	1.21 ± 0.38	0.82 ± 0.20	12.47 ± 0.31
	VAT	81.18 ± 0.69	1.02 ± 0.11	0.72 ± 0.06	11.43 ± 0.96
	Co-training	81.49 ± 0.65	**0.76 ± 0.08**	0.61 ± 0.05	15.03 ± 1.23
	Mean Teacher	84.04 ± 0.44	1.21 ± 0.23	0.79 ± 0.12	8.63 ± 0.79
	AE-prior	81.22 ± 0.64	0.92 ± 0.12	0.71 ± 0.07	13.35 ± 0.70
	CAT (no adv)	80.40 ± 0.65	1.06 ± 0.14	0.73 ± 0.08	14.62 ± 1.57
	CAT	83.77 ± 0.17	1.05 ± 0.06	0.76 ± 0.04	10.25 ± 0.99
	CoT + CAT	82.60 ± 0.20	0.81 ± 0.12	0.60 ± 0.06	13.79 ± 0.45
	MT + CAT	**84.09 ± 0.77**	0.82 ± 0.07	**0.60 ± 0.03**	**7.94 ± 0.85**
3 %	Baseline	71.86 ± 1.50	1.81 ± 0.05	1.23 ± 0.01	15.80 ± 2.14
	Entropy min	72.48 ± 1.80	1.96 ± 0.23	1.33 ± 0.17	15.83 ± 0.67
	VAT	74.83 ± 0.77	1.68 ± 0.18	1.13 ± 0.10	13.44 ± 1.38
	Co-training	76.22 ± 2.33	1.58 ± 0.20	1.06 ± 0.13	19.74 ± 3.30
	Mean Teacher	82.39 ± 0.48	1.13 ± 0.12	0.77 ± 0.06	8.45 ± 0.49
	AE-prior	71.56 ± 0.28	1.91 ± 0.39	1.27 ± 0.19	15.78 ± 1.39
	CAT (no adv)	70.47 ± 0.54	1.65 ± 0.14	1.16 ± 0.08	15.15 ± 1.35
	CAT	80.71 ± 0.20	1.12 ± 0.10	1.12 ± 0.10	8.44 ± 1.71
	CoT + CAT	76.95 ± 1.35	**0.90 ± 0.01**	**0.69 ± 0.03**	18.38 ± 0.82
	MT + CAT	**82.41 ± 0.92**	1.11 ± 0.08	0.77 ± 0.05	**7.97 ± 1.85**

with Mean Teacher achieves the highest DSC and lowest N-conn for all datasets and percentage of labeled examples. To verify the significance of improvements brought by our method, we ran a one-tailed paired t-test for each segmentation task and performance metric. We underline the values with significant improvements ($p < 0.05$) with respect to the second best method. As can be observed, for most cases, our method yields best performance in terms of DSC and significant improvements with respect to constraint satisfaction. We also provide in Figure 3.8 boxplots of performance on ACDC and Hippocampus datasets with 3% labeled examples. Once more, we observe the superiority of the proposed method.

Table 3.11 reports the results for the left ventricle (LV) segmentation task of ACDC with convexity constraints and 3% of training examples being labeled. Once again, CAT outperforms the supervised Baseline with a 2.88% higher DSC and 0.97% fewer pixels of the foreground's

Table 3.8 DSC, MHD, ASSD and non-connectivity (N-conn) for segmenting the PROMISE12. We report the mean and stdev. obtained over three runs.

Label %	Method	DSC (%)	MHD (mm)	ASSD (mm)	N-conn (%)
100 %	Baseline	87.99 ± 0.20	1.19 ± 0.07	1.00 ± 0.05	6.80 ± 0.78
8 %	Baseline	66.79 ± 2.59	2.57 ± 0.15	2.28 ± 0.15	21.72 ± 5.36
	Entropy min	68.68 ± 0.79	2.64 ± 0.19	2.27 ± 0.07	21.42 ± 1.24
	VAT	73.33 ± 0.64	2.90 ± 0.08	2.22 ± 0.04	13.43 ± 0.46
	Co-training	67.64 ± 0.84	2.49 ± 0.10	2.24 ± 0.10	24.52 ± 1.53
	Mean Teacher	79.93 ± 0.34	2.27 ± 0.17	**1.77 ± 0.13**	12.58 ± 2.68
	AE-prior	68.82 ± 0.71	2.63 ± 0.19	2.11 ± 0.12	24.83 ± 0.24
	CAT (no adv)	72.81 ± 1.58	3.19 ± 0.49	2.36 ± 0.21	20.13 ± 7.14
	CAT	75.39 ± 0.88	3.80 ± 0.13	2.58 ± 0.02	12.97 ± 1.83
	CoT + CAT	75.48 ± 0.82	**2.22 ± 0.28**	1.86 ± 0.16	20.09 ± 3.67
	MT + CAT	**80.11 ± 1.13**	2.59 ± 0.22	1.89 ± 0.10	**7.59 ± 1.43**
5 %	Baseline	55.95 ± 1.80	7.95 ± 2.19	5.34 ± 1.21	29.06 ± 2.08
	Entropy min	56.39 ± 3.01	6.05 ± 1.10	4.47 ± 0.79	26.74 ± 2.12
	VAT	62.89 ± 4.20	4.75 ± 0.78	3.78 ± 0.72	17.37 ± 4.80
	Co-training	52.60 ± 0.67	9.78 ± 1.79	6.54 ± 1.05	34.59 ± 2.59
	Mean Teacher	75.95 ± 4.20	3.43 ± 1.62	2.59 ± 1.03	15.30 ± 8.52
	AE-prior	59.51 ± 0.69	4.85 ± 1.27	3.35 ± 0.71	29.14 ± 0.94
	CAT (no adv)	66.63 ± 2.22	4.57 ± 0.56	3.23 ± 0.25	12.25 ± 0.91
	CAT	66.71 ± 1.65	4.80 ± 0.29	3.46 ± 0.18	11.10 ± 0.91
	CoT + CAT	67.75 ± 0.99	5.16 ± 1.18	3.50 ± 0.53	11.23 ± 3.32
	MT + CAT	**77.42 ± 3.53**	**2.99 ± 0.88**	**2.23 ± 0.50**	**10.47 ± 6.76**

convex hull predicted as background. Adding adversarial training also benefits segmentation accuracy and constraint satisfaction in this setting, with a 2.53% higher DSC and 0.57% lower N-conv. The AE-prior approach also improves DSC by 1.63% compared to the supervised Baseline, however it has a worse MHD, ASSD and convexity satisfaction (N-conv) than our CAT method. Last, as observed for the connectivity constraint, using CAT as plug-in method on top of Co-training and Mean Teacher helps obtain more convex predictions with a lower N-conn. Note that this does not necessarily translate into a higher DSC since the LV is not always perfectly convex.

Visualization of results We also confirm the effectiveness of our method by visually comparing segmentation results of tested approaches. Figure 3.9 and Figure 3.10 show some examples for test images of the three datasets. The first five rows in Figure 3.9 give segmentations on

Table 3.9 DSC, MHD, ASSD and non-connectivity (N-conn) for segmenting the Prostate. We report the mean and stdev. obtained over three runs.

Label %	Method	DSC (%)	MHD (mm)	ASSD (mm)	N-conn (%)
100 %	Baseline	70.55 ± 1.85	2.57 ± 0.43	2.08 ± 0.26	20.08 ± 0.25
25 %	Baseline	61.89 ± 2.88	3.51 ± 0.94	2.71 ± 0.51	18.71 ± 2.45
	Entropy min	62.36 ± 0.89	2.94 ± 0.40	2.17 ± 0.16	19.78 ± 3.54
	VAT	62.36 ± 1.94	3.14 ± 0.38	2.41 ± 0.32	19.81 ± 0.07
	Co-training	63.29 ± 0.37	2.52 ± 0.15	2.07 ± 0.16	18.58 ± 1.34
	Mean Teacher	65.99 ± 2.50	2.43 ± 0.73	2.00 ± 0.52	15.82 ± 0.99
	AE-prior	63.05 ± 1.13	2.73 ± 0.12	2.38 ± 0.09	19.57 ± 1.11
	CAT (no adv)	62.29 ± 1.46	3.02 ± 0.24	2.27 ± 0.04	18.18 ± 2.45
	CAT	63.02 ± 0.84	3.61 ± 0.83	2.60 ± 0.59	19.01 ± 1.00
	CoT + CAT	63.67 ± 0.62	2.39 ± 0.29	1.89 ± 0.17	21.58 ± 1.50
	MT + CAT	**66.76 ± 2.71**	**2.35 ± 0.41**	**1.82 ± 0.27**	**14.91 ± 1.26**

Table 3.10 DSC, MHD, ASSD and non-connectivity (N-conn) for segmenting the Hippocampus. We report the mean and stdev. obtained over three runs.

Label %	Method	DSC (%)	MHD (mm)	ASSD (mm)	N-conn (%)
100 %	Baseline	85.90 ± 0.11	0.88 ± 0.01	0.75 ± 0.01	12.47 ± 0.53
3 %	Baseline	79.78 ± 0.43	1.31 ± 0.13	1.07 ± 0.07	14.89 ± 0.46
	Entropy min	80.53 ± 0.20	1.19 ± 0.06	1.00 ± 0.03	14.21 ± 0.67
	VAT	80.47 ± 0.37	1.41 ± 0.28	1.11 ± 0.14	12.80 ± 0.57
	Co-training	81.65 ± 0.16	**1.06 ± 0.01**	0.91 ± 0.01	16.08 ± 0.38
	Mean Teacher	81.78 ± 0.13	1.12 ± 0.02	0.95 ± 0.01	12.62 ± 0.79
	AE-prior	80.45 ± 0.08	1.22 ± 0.09	1.01 ± 0.04	14.62 ± 0.55
	CAT (no adv)	80.00 ± 0.14	1.25 ± 0.11	1.03 ± 0.06	14.69 ± 0.20
	CAT	80.96 ± 0.21	1.26 ± 0.11	0.98 ± 0.11	11.69 ± 0.19
	CoT + CAT	81.66 ± 0.13	1.08 ± 0.02	**0.92 ± 0.01**	15.74 ± 0.02
	MT + CAT	**82.21 ± 0.26**	1.10 ± 0.01	0.94 ± 0.01	**10.65 ± 0.68**

ACDC, PROMISE12, and Prostate datasets. As can be seen, the Baseline and Entropy min method give generally a poor segmentation with unconnected regions. While VAT and AE-prior improve this segmentation, they may still yield unconnected foreground for difficult examples. In comparison, our method produces more plausible segmentations avoiding such unconnected regions. The last two rows of Figure 3.9 show results for the LV segmentation task with convexity

Table 3.11 DSC, MHD, ASSD and non-convexity (N-conv) for segmenting the left ventricle (LV) of ACDC. We report the mean and stdev. obtained over three runs.

Label %	Method	DSC (%)	MHD (mm)	ASSD (mm)	N-conv (%)
100 %	Baseline	94.00 ± 0.09	0.43 ± 0.16	0.33 ± 0.07	4.18 ± 0.42
3 %	Baseline	82.49 ± 1.20	1.90 ± 0.30	1.23 ± 0.13	4.56 ± 0.18
	Entropy min	82.93 ± 0.75	2.64 ± 0.54	1.62 ± 0.28	4.34 ± 0.37
	VAT	83.61 ± 0.22	2.21 ± 0.46	1.43 ± 0.22	4.05 ± 0.47
	Co-training	83.20 ± 0.17	1.81 ± 0.46	1.19 ± 0.23	4.23 ± 0.47
	Mean Teacher	**88.99 ± 0.92**	1.29 ± 0.40	0.81 ± 0.22	4.12 ± 0.35
	AE-prior	84.12 ± 1.10	2.48 ± 0.21	1.54 ± 0.15	4.57 ± 0.32
	CAT (no adv)	82.84 ± 0.86	2.09 ± 0.54	1.37 ± 0.27	4.16 ± 0.36
	CAT	85.37 ± 0.87	**0.81 ± 0.05**	**0.65 ± 0.03**	**3.59 ± 0.11**
	CoT + CAT	84.89 ± 0.39	0.89 ± 0.05	0.72 ± 0.04	3.91 ± 0.08
	MT + CAT	87.21 ± 0.52	1.38 ± 0.19	0.89 ± 0.10	3.61 ± 0.14

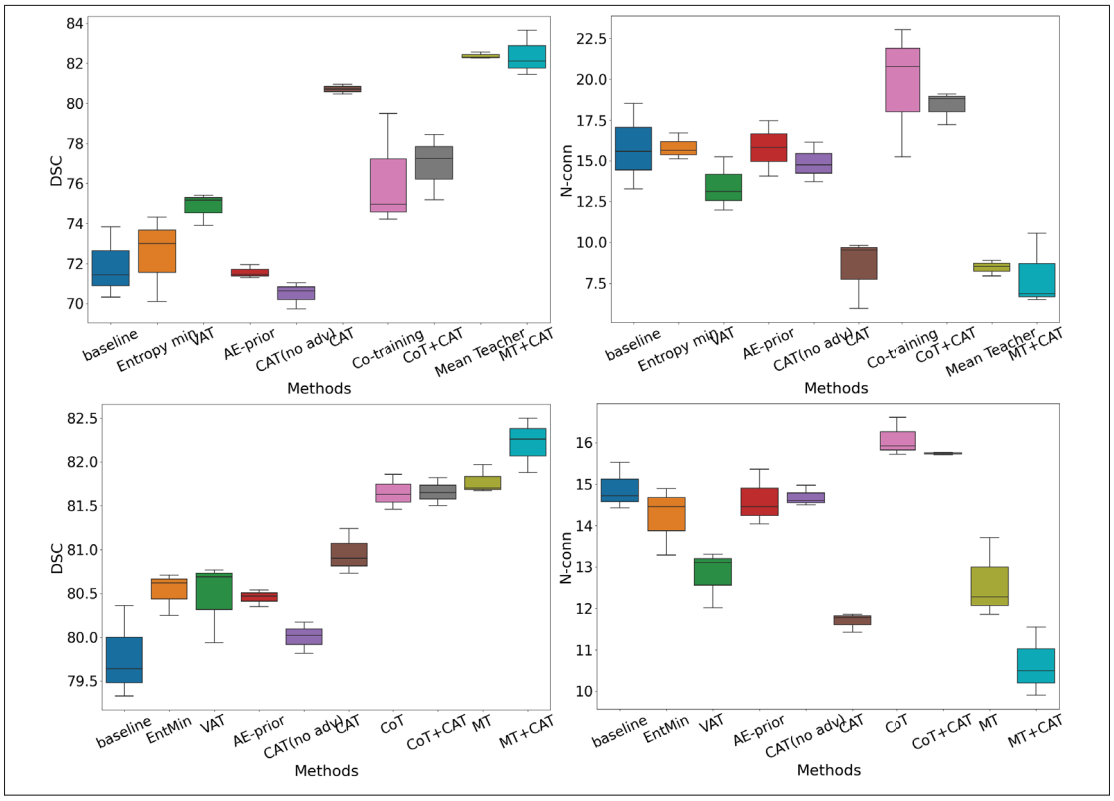


Figure 3.8 Boxplots of performance on ACDC (first row) and Hippocampus (second row) with 3% labeled examples.

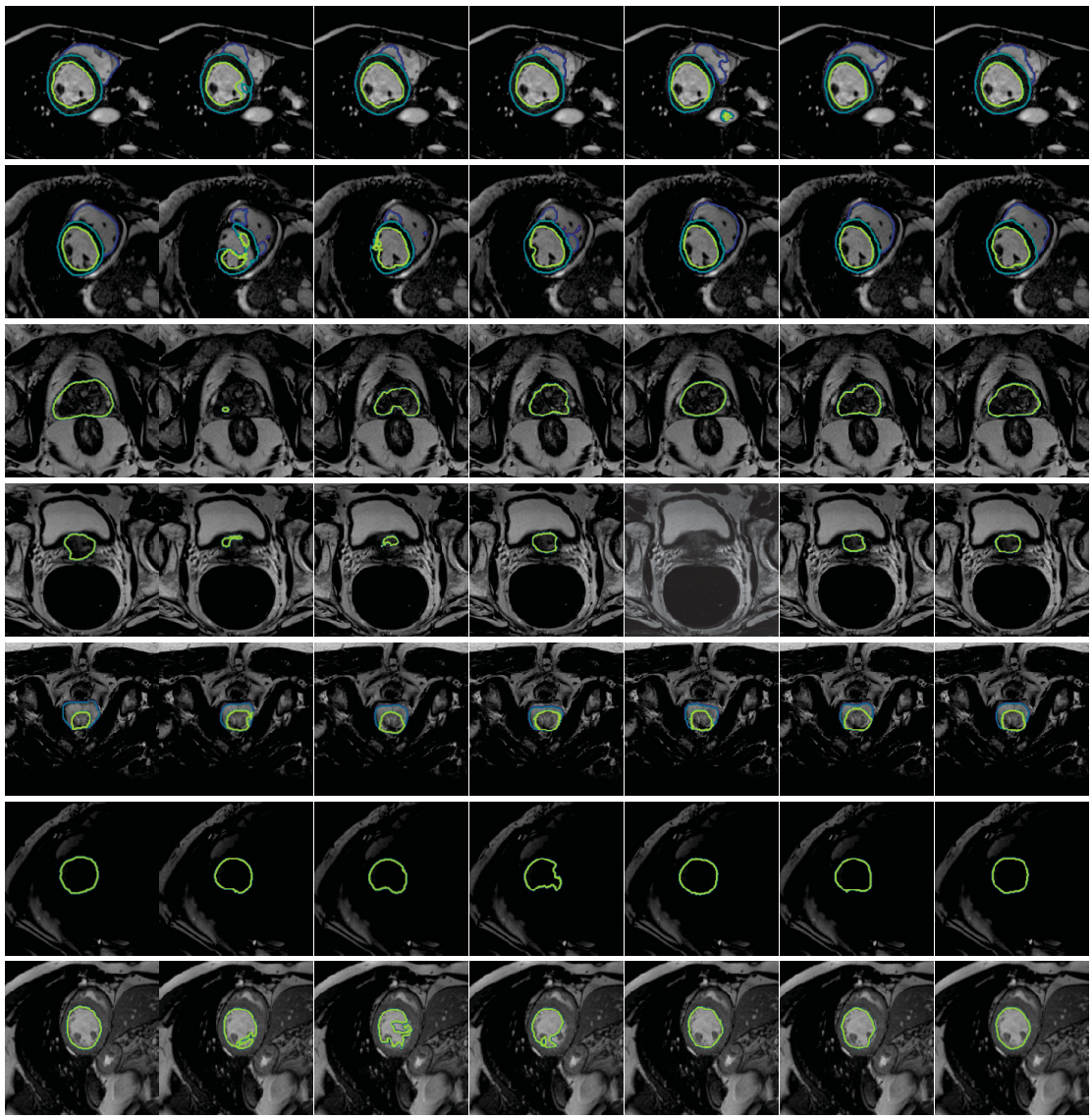


Figure 3.9 Visual results comparison of tested methods. The first two rows show segmentations for connectivity connectivity constraints on ACDC, the middle four rows segmentations of prostate from the Promise12 and Prostate datasets, also with connectivity constraints, and the last two rows segmentations of LV with convexity constraints.The first two rows show segmentations for connectivity connectivity constraints on ACDC, the middle four rows segmentations of prostate from the Promise12 and Prostate datasets, also with connectivity constraints, and the last two rows segmentations of LV with convexity constraints.

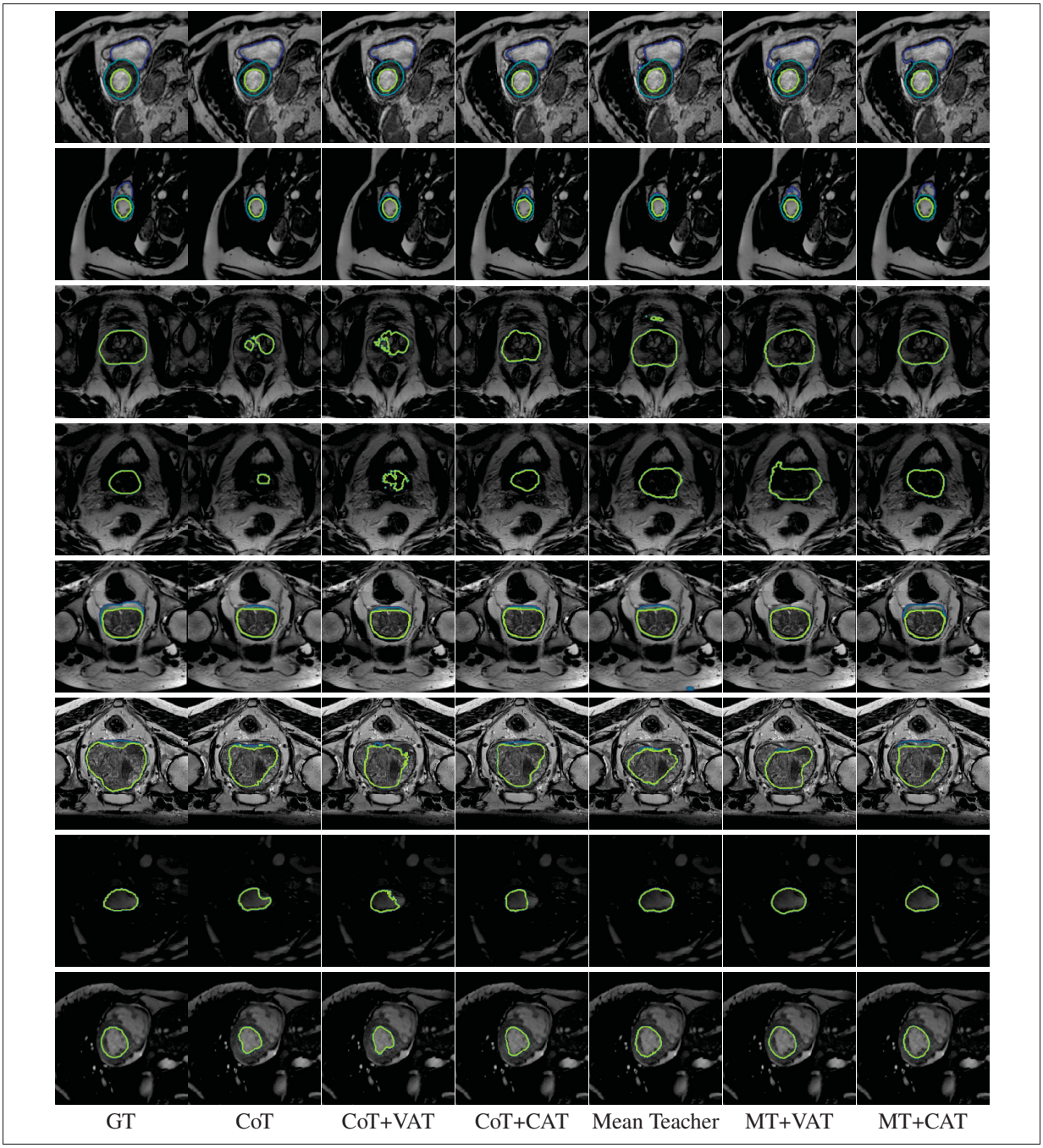


Figure 3.10 Visual results comparison of CAT and VAT plug-in variants on Co-training and Mean Teacher. The first two rows show segmentations for connectivity connectivity constraints on ACDC, the middle four rows segmentations of prostate from the PROMISE12 and Prostate datasets, also with connectivity constraints, and the last two rows segmentations of LV with convexity constraints.

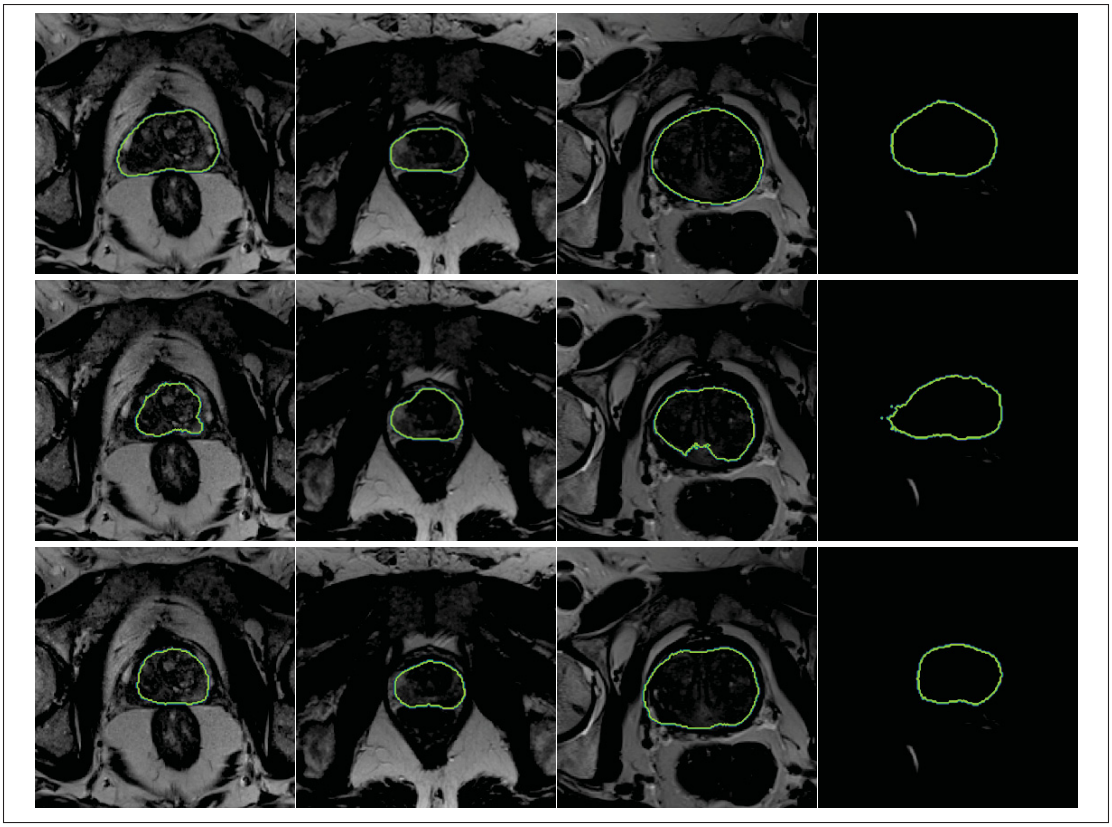


Figure 3.11 Visual results comparison with respect to symmetry. The top row shows the ground truth, the second row shows VAT segmentations, and the bottom row shows CAT segmentations.

constraints. Once again, the proposed method achieves a better, more convex segmentation even for challenging examples with low contrast.

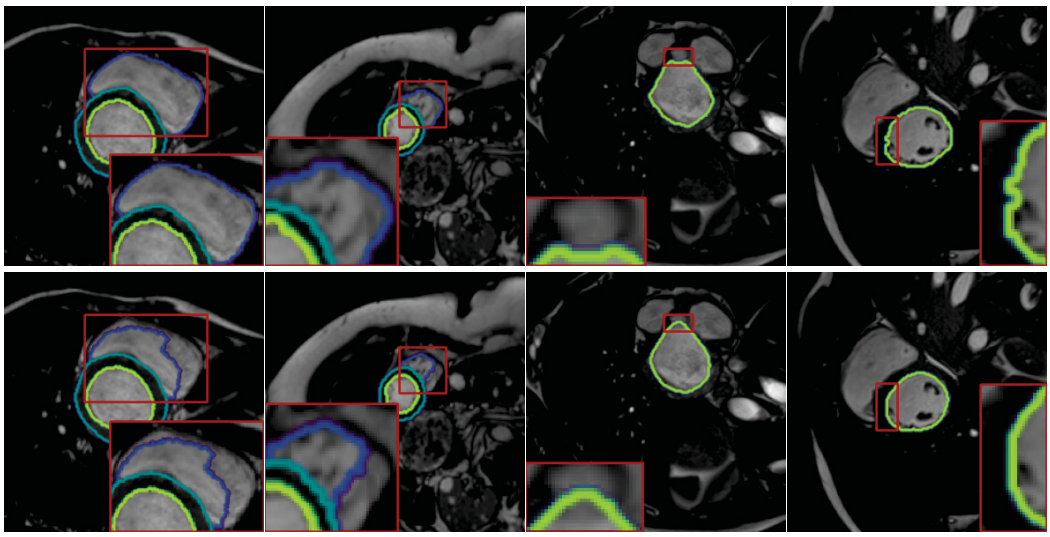
We further validate the advantages of our method as plug-in on top of Co-training and Mean Teacher by visualizing segmentation examples in Figure 3.10. As shown, CoT + CAT and MT + CAT produce a more accurate segmentation than the corresponding VAT variants. We also demonstrate the usefulness of symmetry constraints by conducting experiments on the PROMISE12 dataset where the shape of prostate is approximately symmetric. Figure 3.11 shows visual examples of segmentations obtained by the proposed CAT method and VAT. We see that our method with horizontal symmetry constraints produces more plausible segmentations than VAT which does not enforce this property.

Table 3.12 Training and inference time of the tested methods, for a batch size of 1. The values of CAT(no adv) and CAT represents the training time for connectivity / convexity.

Method	Training time (ms)	Inference time (ms)
Baseline	111	87
Entropy min	125	87
VAT	250	87
Co-training	233	128
MT	152	87
CAT (no adv)	167 / 158	87
CAT	270 / 263	87

3.5.3 Computational efficiency

To demonstrate our method's efficiency, we analyze the average training time and inference time of tested approaches using a batch size of 1, for the segmentation with connectivity and convexity constraints. All methods were implemented in Pytorch and were run on an Nvidia 3070Ti GPU. Results of this analysis are summarized in Table 3.12. The baseline model, which only needs to compute a single loss per pass, has the lowest training times. Since it has to generate adversarial examples, VAT incurs a longer training time. Co-training needs to train two separate models, however the parameters of these two models can be updated in parallel. Although Mean Teacher also uses two networks in training, only the Student model is updated via back-propagation (the teacher's parameters are updated as an exponential moving average of the student's). CAT (no adv), which imposes the constraint directly on original unlabeled data without generating adversarial examples, has a training time 1.5× longer than the baseline (158-167 ms vs 111 ms) and comparable to VAT. As it needs to generate adversarial examples and compute the constraint loss, our CAT method has training times 2.5× larger than the baseline. However, as these steps are only performed during training, our method does not incur additional memory or computational cost at inference time.



(a) Conn (b) Conn (c) Conv (d) Conv

Figure 3.12 Failure cases of the proposed method. The first row shows the ground truth. The second row shows the failed segmentation produced by our method. (a)–(b) are two examples of failed case with connectivity (Conn), and (c)–(d) are two examples of failed case with convexity (Conv).

3.5.4 Discussion

Our constrained adversarial training (CAT) method for segmentation extends VAT by generating adversarial examples that maximizes both output divergence and constraint violation. This is achieved with an efficient optimization strategy based on the REINFORCE algorithm that can obtain useful gradients for non-differentiable constraints. Our CAT method outperforms other compared approaches in most of the cases, generating segmentations that better satisfy a given set of anatomical constraints. Nevertheless, it may also have some failure cases. For connectivity, we penalize the foreground pixels which are not connected to the main region (containing the seed of the flood-fill algorithm). The model is therefore encouraged to remove isolated regions whether they are part of the target structure or not, which can result in an underestimated foreground. Conversely, for convexity constraints, we penalize background pixels which are inside the convex hull of the foreground, hence we enforce the model to expand the foreground until it becomes convex. This may lead to an over-enlarged segmentation. Figure 3.12 gives failure examples corresponding to these scenarios: an underestimated segmentation of the RV

when using connectivity constraints, and an enlarged segmentation of the LV for the convexity case.

3.6 Conclusion

We proposed CAT, a semi-supervised method for segmentation that can incorporate complex anatomical constraints into a deep neural network to produce more plausible predictions. Our method exploits unlabeled examples in an adversarial training strategy that regularizes the network and helps it learn constraints. By making the network robust to adversarial perturbations that maximize both prediction divergence and constraint violation, we improve the robustness of our method and an obtain useful gradients for learning complex constraints on the fly.

Our ablation study on synthetic data demonstrated the usefulness of our method's Soft satisfaction strategy and Reverse reward formulation, which improve both segmentation accuracy and connectivity/convexity on test examples. It also showed the positive impact of the adversarially-trained constraint loss when increasing its weights in the overall objective function. Moreover, results on three benchmark datasets related to cardiac and prostate segmentation revealed the superior performance of our method compared to state-of-art approaches. As a stand-alone method, on the lowest supervision settings of the ACDC and PROMISE12 datasets, CAT gave a 8.85–10.76% higher DSC than the Baseline model while also increasing pixel foreground connectivity by 7.36–17.96%. Likewise, adding our method as plug-in on top of Co-training and Mean Teacher improved both accuracy and constraint satisfaction in almost all test cases. As future work, we plan to explore a broader range of segmentation constraints.

CHAPTER 4

SHAPE-AWARE JOINT DISTRIBUTION ALIGNMENT FOR CROSS-DOMAIN IMAGE SEGMENTATION

Ping Wang[1] , Jizong Peng[1] , Marco Pedersoli[1] , Yuanfeng Zhou[2] , Caiming Zhang[2] , Christian Desrosiers[1]

[1] Department of Software and IT Engineering, École de Technologie Supérieure, 1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3

[2] School of Software, Shandong University, 1500 Middle of Shunhua Road, Jinan, Shan Dong, China 250101

Article published in Journal «IEEE Transaction on Medical Imaging» in February 2023.

4.1 Introduction

In recent years, deep learning models such as convolutional neural networks (Ronneberger *et al.*, 2015) and transformers (Dai, Gao & Liu, 2021b) have led to unprecedented advancements in semantic segmentation. However, the use of such models in clinical practice remains limited due to two major challenges: the scarcity of labeled data for training deep neural networks with millions of parameters, and the cross-site variability of data stemming from differences in the demographics of imaged subjects, imaging modalities or even equipment. The first challenge is typically addressed via semi-supervised learning methods (Cheplygina *et al.*, 2019) that exploit the greater abundance of unlabeled images, in addition to the few available annotated images. On the other hand, the problem of data variability is the focus of *domain adaptation* (DA) approaches, which seek to make a model trained on data from a *source* domain perform well on examples from a new *target* domain with limited or no labeled data from this new domain. In the most difficult setting, called *unsupervised domain adaptation* (UDA), labeled images are only provided for the source domain. Figure 4.1 shows domain shift and domain-invariant information across domains (MR and CT).

A broad range of deep learning DA approaches have been proposed for medical image segmentation (Guan & Liu, 2021), including techniques based on adversarial learning (Goodfellow *et al.*,

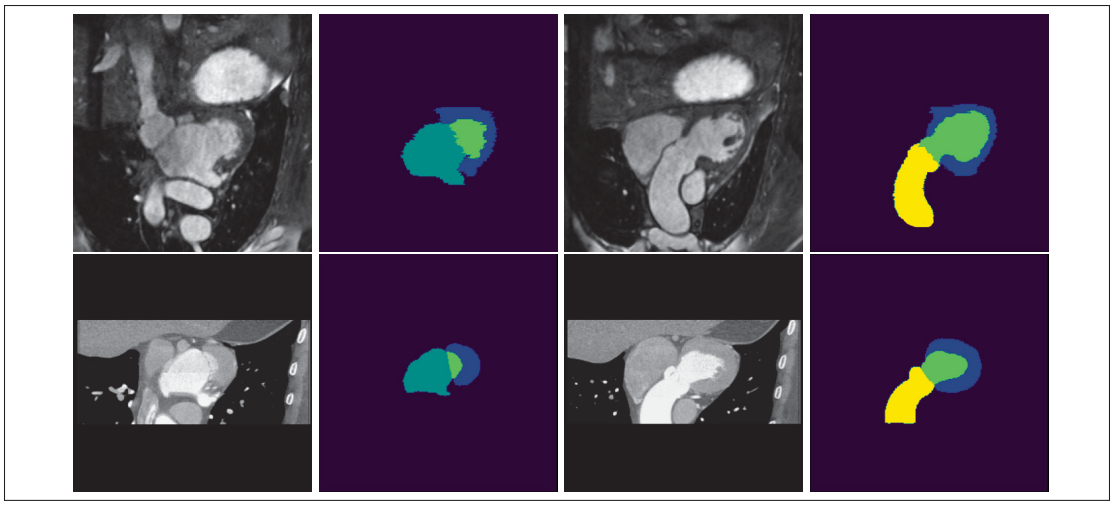


Figure 4.1 Illustration of cross domain shift and domain-invariant spatial relationships on cardiac data. The first row shows the MR images and corresponding annotations, and the second row shows the CT images and corresponding annotations. Images of MR and CT, which have similar annotations, are different in data distribution, that corresponds to a domain shift. Though with domain shift, the annotations for tissues across domains are inherently same, with same number of classes and same spatial relationship between classes.

2020), pseudo-labeling (Xia *et al.*, 2020b; Wang, Peng & Zhang, 2021b), entropy minimization (Vu *et al.*, 2019) and uncertainty estimation (Yu *et al.*, 2019). In a recent work (Bateson *et al.*, 2020), domain-agnostic constraints were used to adapt a network for cross-modality image segmentation. Specifically, the proposed method minimizes the KL divergence between the class marginal distribution of the network for target examples and a reference empirical distribution estimated on source examples. Combined with an entropy minimization loss, this method guides the segmentation network to generate confident predictions for target images which are globally similar to those for source images, without actually needing the source images during adaptation. While it was shown to outperform approaches based on adversarial learning and pseudo-labeling, this method suffers from two limitations. First, it only aligns class-level statistics (i.e., relative size of classes) between domains and, thus, does not fully exploit the spatial relationships between classes in the images. Secondly, it requires image-level tags indicating the presence

or absence of a given class in each image to work, hence it cannot be used in an unsupervised domain adaptation scenario.

In this paper, we propose a novel UDA approach for cross-domain image segmentation based on the idea of shape-aware joint distribution alignment. Our approach computes the joint class distribution between the prediction for two pixels whose relative position corresponds to a certain displacement vector. To account for variability in the size and spatial relationships of regions to segment, we compute this joint matrix for multiple displacements in a given set. Cross-domain adaptation is then achieved by aligning the joint distributions computed for source and target examples. The main contributions of our work are as follows:

- We improve the constraint-based DA method of (Bateson *et al.*, 2020) by incorporating high-order statistics that measure the joint distribution of classes at relative positions corresponding to different orientations and distances;
- We propose an efficient multi-scale strategy that encodes long range relationships between classes in the learned statistics;
- We further extend this approach to feature maps in intermediate layers of the network by computing the cross-correlation between them;
- We demonstrate the advantage of our approach on the cross-modality segmentation of cardiac structures and prostate regions, showing its better performance compared to state-of-art methods for this task. The code is available at https://github.com/WangPing521/Domain_adaptation_shape_prior.

The next section presents related work on unsupervised domain adaptation. The proposed method is then detailed in Section 4.3. Experiments to evaluate this method are described in Section 4.4, and results presented in Section 4.5.

4.2 Related work

Unsupervised domain adaptation for medical image segmentation has generated a growing interest in the last years. One of the earliest approaches to address the domain-shift issue, instance

weighting (Wachinger, Reuter, Initiative *et al.*, 2016; Goetz *et al.*, 2015), identifies training examples similar to target ones and gives these examples a higher importance during model optimization. Recent methods based on deep learning further reduce the distribution discrepancy between domains using two main strategies: 1) aligning the marginal statistics between the source and target domains; 2) constructing an intermediate representation that embeds domain-agnostic knowledge shared by both domains. Several works (Kumagai & Iwata, 2019; Wang *et al.*, 2020) align intermediate features across domains by minimizing their maximum mean discrepancy (MMD). In (Ganin *et al.*, 2016), adversarial learning is used to train a neural network so that the intermediate feature maps (Dou *et al.*, 2019) or predicted outputs (Tsai *et al.*, 2018; Tsai, Sohn, Schulter & Chandraker, 2019) of source and target examples are indistinguishable to a domain classifier. Another popular approach, based on style transfer, uses generative adversarial networks (GANs) to alter the appearance of input images from one domain to the other while preserving their semantic structures (Chen *et al.*, 2020; Ouyang, Kamnitsas, Biffi, Duan & Rueckert, 2019; Yang *et al.*, 2019b; Zhao, Xu, Li, Zeng & Guan, 2021; Vesal, Gu, Kosti, Maier & Ravikumar, 2021). Despite their success in several cross-domain segmentation settings, these adversarial methods need to solve a challenging minimax optimization problem which makes their training unstable (Zhang, Yang & Zheng, 2018; Wu & Zhuang, 2020) and often leads to mode collapse (Liu, Tang, Zhou & Qiu, 2019a).

As simple alternatives to adversarial learning, DA approaches based on entropy minimization (Vu *et al.*, 2019; Li *et al.*, 2021) and pseudo-labeling (Xia *et al.*, 2020b; Wu, Chen, Xiong, Chen & Sun, 2021) have attracted growing attention. Ill-adapted models often provide unconfident and unrealistic predictions for images from a new domain. Entropy minimization seeks to increase the prediction confidence for target examples by minimizing the entropy of the output distribution. In a recent work (Bateson *et al.*, 2020), this approach is enhanced with a domain-invariant prior that enforces the relative class-wise distribution of target examples to be same as a reference source distribution. Unlike our method, this approach requires image-level tags for target examples and does not exploit high-order statistics encoding spatial relationships between classes. Similarly, Bateson *et al.* (Bateson, Dolz, Kervadec, Lombaert & Ayed, 2021) estimated

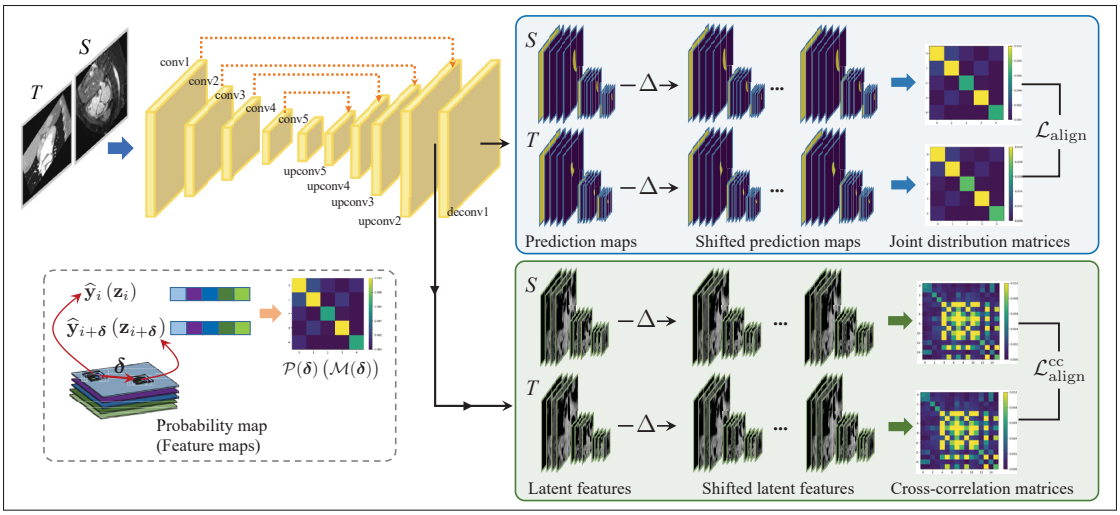


Figure 4.2 Schematic diagram of our proposed information invariant alignment method for unsupervised domain adaptation. Apart from utilizing a supervised loss on the source domain, our method proposes a shape-aware information invariant alignment loss, i.e. the alignment loss of joint probability distributions from the predicted classes and the alignment loss of cross-correlation matrix from high-level latent layer. The combination contributes to improve the inherent semantic segmentation despite the domain shift. The left bottom figure shows a joint matrix (cross-correlation matrix) estimation with a displacement.

the relative size of segmentation classes using a regression network, and then imposed the size predicted for target images to be within a pre-defined range. However, it requires solving a difficult size prediction task and having reliable size bounds for the different classes. In contrast, pseudo-labeling methods identify confident predictions for target examples, using prediction uncertainty (Wu *et al.*, 2021) or the disagreement between multiple models (Xia *et al.*, 2020b), and then consider these predictions as labels in a standard supervised training loss. While this approach works well when the domain gap is small, it generally collapses for larger gaps since the pseudo-labels are then very noisy.

4.3 The proposed method

We start by defining the unsupervised domain adaptation (UDA) problem considered in our work. Let $\mathcal{S} = \{(\mathbf{x}_s, \mathbf{y}_s)\}_{s=1}^{N_S}$ be the set of labeled source examples and $\mathcal{T} = \{\mathbf{x}_t\}_{t=1}^{N_T}$ the unlabeled

examples from the target domain. Here, $\mathbf{x}_s, \mathbf{x}_t \in \mathbb{R}^{W \times H}$ are 2D images of size $W \times H$ and $\mathbf{y}_s \in \{0,1\}^{W \times H \times C}$ is the corresponding ground-truth segmentation for C classes. The goal is to learn a segmentation network $f_\theta(\cdot)$, using $\mathcal{S}$ and $\mathcal{T}$ as training data, which can map a target image $\mathbf{x}_t$ to its corresponding segmentation map $\mathbf{y}_t$.

Our proposed cross-domain segmentation framework is illustrated in Figure 4.2. Unlike adversarial learning or knowledge distillation methods, which also require a discriminator or a teacher network, our framework is composed of a single segmentation network. The network is trained with labeled source examples and unlabeled target images by minimizing a loss function combining three learning objectives:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{sup}} + \lambda_{\text{ent}}\mathcal{L}_{\text{ent}} + \lambda_{\text{align}}\mathcal{L}_{\text{align}}. \tag{4.1}$$

In this loss, $\mathcal{L}_{\text{sup}}$ is a standard supervised loss using the labeled source data $\mathcal{S}$. For this paper we use the well-known cross-entropy segmentation loss:

$$\mathcal{L}_{\text{sup}} = \frac{1}{N_S} \sum_{s=1}^{N_S} \mathcal{H}\big(\mathbf{y}_s, f_\theta(\mathbf{x}_s)\big) \tag{4.2}$$

where $\mathcal{H}(\cdot,\cdot)$ is the mean cross-entropy over pixels, computed as

$$\mathcal{H}\big(\mathbf{y}, \widehat{\mathbf{y}}\big) = -\frac{1}{W \times H} \sum_{i=1}^{W \times H} \sum_{k=1}^{C} y_{ik} \log \widehat{y}_{ik}. \tag{4.3}$$

On the other hand, $\mathcal{L}_{\text{ent}}$ is the entropy minimization loss used to enforce high confidence in target domain $\mathcal{T}$. It is defined as

$$\mathcal{L}_{\text{ent}} = \frac{1}{N_T} \sum_{t=1}^{N_T} \mathcal{H}\big(f_\theta(\mathbf{x}_t), f_\theta(\mathbf{x}_t)\big). \tag{4.4}$$

The domain adaptation loss $\mathcal{L}_{\text{align}}$, which uses both labeled source and unlabeled target data, leverages high-order statistics that encode spatial relationships between classes and enforces these statistics to be consistent across domains. Specifically, it computes the joint class distribution

between pixels at different relative positions, for both source and target images, and then minimizes the discrepancy between these distributions. The relative importance of these learning objectives is controlled by hyper-parameters λ_{ent} and λ_{align}.

4.3.1 Shape-aware joint distribution alignment loss

We use high-order statistics on the segmentation predictions as a domain-invariant prior for adapting the network to target domain data. The statistics measure the joint probability $\mathcal{P}(\boldsymbol{\delta})$ of predicting specific classes at two pixels $i = (w, h)$ and $i' = (w', h')$ whose relative position corresponds to a 2D pixel displacement $\boldsymbol{\delta} = (\delta_w, \delta_h)$, i.e. $w' = w + \delta_w$ and $h' = h + \delta_h$.

For the same class k, joint probability $\mathcal{P}_{kk}(\boldsymbol{\delta})$ encodes information on the size of the region corresponding to this class. For instance, a region whose maximum length is $L_{\max}$ pixels should have a joint probability of $\mathcal{P}_{kk}(\boldsymbol{\delta}) = 0$ for displacements such that $\|\boldsymbol{\delta}\| > L_{\max}$. Likewise, the joint probability $\mathcal{P}_{jk}(\boldsymbol{\delta})$ for two different classes j and k models the spatial relationship of their corresponding regions. Hence, regions far from each other should have a low $\mathcal{P}_{jk}(\boldsymbol{\delta})$ for any $\boldsymbol{\delta}$, whereas nearby ones that follow a certain spatial relationship (e.g., one is always above the other) should have a high $\mathcal{P}_{jk}(\boldsymbol{\delta})$ for displacements $\boldsymbol{\delta}$ corresponding to this relationship.

Let $\widehat{\mathbf{y}} = f_\theta(\mathbf{x}) \in [0, 1]^{W \times H \times C}$ be the predicted class probabilities for a given image $\mathbf{x}$ and define as $i + \boldsymbol{\delta}$ the pixel index corresponding to a displacement $\boldsymbol{\delta}$ from index i. We estimate the joint probability for classes j and k on a batch of examples $\mathcal{B}$, as follows:

$$\mathcal{P}_{jk}(\boldsymbol{\delta}) \;=\; \frac{1}{|\mathcal{B}|} \sum_{\mathbf{x} \in \mathcal{B}} \frac{1}{W \times H} \sum_{i=1}^{W \times H} \widehat{y}_{i,j}\, \widehat{y}_{i+\boldsymbol{\delta},k}. \tag{4.5}$$

This can be expressed in a more compact form using the vector outer product:

$$\mathcal{P}(\boldsymbol{\delta}) \;=\; \frac{1}{|\mathcal{B}|} \sum_{\mathbf{x} \in \mathcal{B}} \frac{1}{W \times H} \sum_{i=1}^{W \times H} \widehat{\mathbf{y}}_i \cdot \widehat{\mathbf{y}}_{i+\boldsymbol{\delta}}^{\top}. \tag{4.6}$$

Denoting as $\mathcal{P}^S(\boldsymbol{\delta})$ and $\mathcal{P}^T(\boldsymbol{\delta})$ the joint distribution matrices for source and target examples, respectively, the proposed DA loss measures the L_1 distance between the source and target matrices for displacements in a set Δ:

$$\mathcal{L}_{\text{align}} = \frac{1}{|\Delta|} \sum_{\boldsymbol{\delta} \in \Delta} \sum_{j,k=1}^{C} \left| \mathcal{P}^S_{jk}(\boldsymbol{\delta}) - \mathcal{P}^T_{jk}(\boldsymbol{\delta}) \right|. \tag{4.7}$$

Note that, following Pinsker's inequality, the L_1 (total variation) distance is related to KL divergence as follows: $\|P - Q\|_1 \leq \sqrt{2 D_{\text{KL}}(P \,\|\, Q)}$. Unlike KL divergence, the L_1 distance has the useful properties of being symmetric and bounded. In our experiments, we found L_1 to be more stable than the latter, due in part to the fact that KL divergence has vanishing gradients when the compared distributions are very different (Arjovsky, Chintala & Bottou, 2017).

One can show that, for a zero displacement $\boldsymbol{\delta}_0 = (0, 0)$, our loss is related to the DA approach in (Bateson *et al.*, 2020) which imposes the marginal class distribution to be the same for source and target images. In this case, elements in the joint matrix are given by

$$\mathcal{P}_{jk}(\boldsymbol{\delta}_0) = \frac{1}{|\mathcal{B}|} \sum_{\mathbf{x} \in \mathcal{B}} \frac{1}{W \times H} \sum_{i=1}^{W \times H} \widehat{y}_{ij} \, \widehat{y}_{ik}. \tag{4.8}$$

If the network is well-trained on source images $\mathbf{x}$, it will have low entropy prediction for these examples, therefore $\mathbf{p}$ will be near binary. We will then have that $\widehat{y}_{ij} \cdot \widehat{y}_{ik} \approx 0$ for $j \neq k$ and $\widehat{y}_{ik} \cdot \widehat{y}_{ik} \approx p_{ik}$. Hence, $\mathcal{P}(\boldsymbol{\delta}_0)$ be a diagonal matrix whose diagonal elements are the estimated class marginals.

4.3.2 Multi-scale joint distribution alignment

A problem with the alignment loss in Eq. (4.7) is that it requires a large number of displacements to model all the possible orientations and distances. While it is possible to compute displacements on a sparse grid, this leads to a worse estimation of statistics. Moreover, we need to compute joint matrices for both domains with sets of displacements in a high resolution, leading to a high computation cost. Instead, we propose a multi-scale strategy which down-scales the

predicted segmentation maps to different spatial resolutions by cascading 2×2 average pooling operations. We then use a small set of displacements Δ on the predictions at each resolution. This strategy has two significant advantages. First, it enables modeling multiple orientations and distances with fewer displacements, i.e., $|\Delta| \times S$ where S is the number of scales. Second, since down-sampling averages the predictions for multiple pixels, it provides a better estimation of the joint matrix compared to sparse grid sampling.

4.3.3 Cross-correlation matrix alignment on latent features

The core idea of the proposed DA method is to align domain-invariant information. In the previous section, we align the joint class distribution at the output of the network. However, semantic information captured by the feature maps of intermediate layers could also benefit from such alignment. Based on this idea, we extend our alignment loss to these intermediate layers.

A simple way to achieve this goal is to project feature vectors to a discrete space of K clusters, for example using a 1×1 convolution followed by a K-way softmax, and then compute the joint on these clusters. However, as shown in our experiments (see Figure 4.5), it is challenging for the model to learn clusters that generalize across different domains, without explicit supervision. Instead, we propose a strategy based on cross-correlation to align the latent features. Let $\mathbf{z} = h(\mathbf{x}) \in \mathbb{R}^{W' \times H'}$ be the feature map at a given layer (typically in the decoder of the segmentation network) encoding high-level semantic information. The cross-correlation matrix $\mathcal{M}(\boldsymbol{\delta})$ for a displacement $\boldsymbol{\delta}$ is estimated from a batch $\mathcal{B}$ as

$$\mathcal{M}(\boldsymbol{\delta}) = \frac{1}{|\mathcal{B}|} \sum_{\mathbf{x} \in \mathcal{B}} \frac{1}{W' \times H'} \sum_{i=1}^{W' \times H'} \mathbf{z}_i \cdot \mathbf{z}_{i+\boldsymbol{\delta}}^{\top}. \tag{4.9}$$

Following the same approach as for the joint distribution, we align the cross-correlation matrices of the source ($\mathcal{M}^S$) and target domains ($\mathcal{M}^T$) for displacements in a set Δ by minimizing the

mean L_1 distance:

$$\mathcal{L}_{\text{align}}^{\text{cc}} = \frac{1}{|\Delta|} \sum_{\delta \in \Delta} \sum_{j,k=1}^{C} \left| \mathcal{M}_{jk}^{S}(\delta) - \mathcal{M}_{jk}^{T}(\delta) \right|. \tag{4.10}$$

Combining this new loss term with the original loss of Eq. (4.1), we get our final total loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{sup}} + \lambda_{\text{ent}} \mathcal{L}_{\text{ent}} + \lambda_{\text{align}} \mathcal{L}_{\text{align}} + \lambda_{\text{align}}^{\text{cc}} \mathcal{L}_{\text{align}}^{\text{cc}}. \tag{4.11}$$

The next section presents the experiment setup used to evaluate our method.

4.4 Experimental setup

4.4.1 Datasets

We test our method on a cardiac segmentation task using the MICCAI 2017 Multi-Modality Whole Heart Segmentation (MMWHS) Challenge (Zhuang *et al.*, 2019) dataset, and on a prostate segmentation task using two datasets, the Prostate MR Image Segmentation (Promise2012) Challenge dataset (Litjens *et al.*, 2014) and Prostate sub-task dataset of the Medical Segmentation Decathlon Challenge (Antonelli *et al.*, 2022) (Prostate). Theses segmentation tasks and corresponding datasets are described below.

Cardiac segmentation: The MMWHS dataset consists of unpaired 20 CT and 20 MRI volumes with ground truth mask from different patients and sites. We employ MRI images as source domain and CT images as target domain. Following (Dou *et al.*, 2020), we choose four main structures for the segmentation task: left ventricle myocardium (LVM), left atrium blood cavity (LAC), left ventricle blood cavity (LVC) and ascending aorta (AA). Images from both modalities are re-sampled to an identical voxel-spacing of 2.0×1.0×1.0 mm^3. We divide volumetric scans into 2D slices along coronal axis and center crop images to 256×256 pixels. Target data are split into train and test sets containing 16 and 4 volumes, respectively. We use 4-fold cross validation on the training set to determine the hyper-parameters of all compared models.

Prostate segmentation: The PROMISE12 dataset consists of multi-centric transversal T2-weighted MR volumes from 50 subjects. Image resolution ranges from 15×256×256 to 54×512×512 voxels with a spacing ranging from 2×0.27×0.27 to 4×0.75×0.75 mm^3. A single region is labeled in the ground-truth. On the other hand, the PROSTATE dataset consists of 48 multi-parametric MRI (32 MRIs are labeled) provided by Radboud University. Two structures are labeled in the ground-truth: peripheral zone (PZ) and central gland (CG). We first merge these two regions into a single one to align with the PROMISE12 dataset. We then employ PROSTATE as source domain and PROMISE12 as target domain. Images from both datasets are re-sampled to an identical voxel-spacing of 2.0×0.5×0.5 mm^3. We slice volumetric scans into 2D slices and center crop obtained images to 256×256 pixels. Target images are split with 80% scans for training, 20% scans for testing. As before, we use 4-fold cross validation on the training set to select the hyper-parameters of all models, within a grid search.

During training, we employ a rich set of transformations as image augmentation, including various affine transformations and gamma correction-based intensity distortion. For all experiments, we report 3D Dice Similarity Coefficient (DSC) as the performance metric for the compared methods and ablation variants. Let S and G be the predicted and ground-truth mask for a given class. The 3D DSC for this class is computed as

$$\mathrm{DSC}(S, G) = \frac{2|S \cap G|}{|S| + |G|}. \tag{4.12}$$

4.4.2 Implementation details

We use UNet (Ronneberger *et al.*, 2015) as our segmentation architecture for all experiments. As suggested in (Boutillon, Conze, Pons, Burdin & Borotikar, 2021; Dou *et al.*, 2020), we replaced the shared batch normalization (BN) layer by domain-specific BNs. The network is trained using a rectified Adam optimizer (Liu *et al.*, 2019b) with a learning rate decay strategy where the initial learning rate 1×10^{-5} is increased by 300 times in the first 10 epochs, followed by a cosine decay policy (Loshchilov & Hutter, 2016).

4.4.3 Compared methods

We compare our method with baselines and six state-of-the-art UDA approaches: EntDA Vu *et al.* (2019), PLDA (Wu *et al.*, 2021), Ent + prior (Bateson *et al.*, 2020), SIFA (Chen *et al.*, 2020), MT-UDA (Zhao *et al.*, 2021) and PointcloudDA (Vesal *et al.*, 2021).

Upper baseline: This supervised baseline is both trained and tested on target (CT) images. It is used to measure the highest performance that can be achieved when no domain shift is present.

Lower baseline: In this second baseline, we train the segmentation network *only* on source (MR) images to estimate its ability to generalizing to target (CT) images.

EntDA (Vu *et al.*, 2019): This method seeks to increase the network's confidence for target examples by minimizing their prediction entropy. The loss function is formulated as $\ell_{\text{total}} = \ell_{\text{sup}} + \lambda_{\text{ent}}\ell_{\text{ent}}$, where ℓ_{ent} is the prediction entropy for target examples as defined in Eq. (4.4). The method corresponds to an ablation variant of our approach, where no alignment loss is used ($\lambda_{\text{align}} = 0$).

PLDA (Wu *et al.*, 2021): PLDA uses the predictions by a source-trained model on target examples as pseudo labels. Unadapted models often output unrealistic and unconfident predictions for target images. This method employs a uncertainty estimation approach based on Monte-Carlo dropout sampling to qualify the reliability of these pseudo labels. Adaptation is achieved by training a second model with rectified pseudo labels on target images, using a cross-entropy loss. However, this method gives poor results when the domain shift is too large since pseudo labels are then unusable. To have a competitive performance, instead of generating pseudo labels from a model trained only on source data (i.e., the *Lower baseline*), we generate pseudo labels from the stronger *EntDA* model.

Ent+prior (Bateson *et al.*, 2020): This method enhances *EntDA* with a KL divergence loss aligning the class marginals (i.e., relative proportion of pixels in each class) of the source and target domains. Unlike (Bateson *et al.*, 2020), which initializes the target domain model weights using a source pre-trained model and leverages image tags during the target DA stage, we

optimize this DA loss jointly with our supervised loss on source images and use no image tags for target images. Specifically, the loss function for this method is defined as

$$\ell_{\text{total}} = \ell_{\text{sup}} + \lambda_{\text{ent}}\ell_{\text{ent}} + \lambda_{\text{KL}}D_{\text{KL}}\big(\mathbf{p}_T \,\|\, \mathbf{p}_S\big) \tag{4.13}$$

where

$$\mathbf{p}_T = \frac{1}{N_T \times W \times H} \sum_{t=1}^{N_T} \sum_{i=1}^{W \times H} [f_\theta(\mathbf{x}_t)]_i \tag{4.14}$$

is the marginal estimated from predictions on target images and $\mathbf{p}_S$ is the *fixed* reference marginal computed from all source images.

SIFA (Chen *et al.*, 2020): The Synergistic Image and Feature Alignment (SIFA) method performs domain alignment at the level of image and feature. At the image level, a CycleGAN (Zhu, Park, Isola & Efros, 2017) composed of a cycle-consistency loss and two discriminators is used to convert labeled source images to the target domain, without the need for paired images. Synthesized target images, along their source ground-truth, are used to train a segmentation network whose encoder is shared with the CycleGAN. On the other hand, feature alignment is performed with a discriminator on the segmentation output, which tries to determine if a given segmentation prediction comes for a real or generated target image. To align internal features of the network, segmentation decoders taking these features as input are added to the model. These features are aligned via an adversarial loss on the auxiliary predictions of added decoders.

MT-UDA (Zhao *et al.*, 2021): This UDA approach exploits a framework based on Mean Teacher, which is composed of a student model and two domain-specific teachers. A dual cycle alignment module (DCAM) based on adversarial learning synthesizes source-like domain images and target-like domain images, each used to train a different teacher. The student model distills the intra-domain knowledge with a loss encouraging prediction consistency with the source-like domain teacher. Similarly, inter-domain knowledge is exploited by enforcing the structural consistency between the student and target-domain teacher.

PointcloudDA (Vesal *et al.*, 2021): This method achieves domain adaptation from three separate components, output space alignment, entropy minimization and point-cloud shape alignment, each one using a different discriminator. For output space alignment, a discriminator is trained in an adversarial manner to recognize if the output of the segmentation network is for a source or a target image, encouraging this network to produce similar predictions across domains. Entropy-driven adversarial learning is also adopted to encourage structural consistency, by enforcing the output uncertainty maps encoded by entropy to be similar across domains. To further improve performance, shape information encoded by a point cloud is aligned between the two domains. Toward this goal, a point cloud regression head attached to the encoder is used to predict a set of points located on the combined surface of the structures to segment. A discriminator based on PointNet, which is trained with an adversarial objective, is then employed to make the predicted point clouds for different domains to be similar.

4.5 Results

To evaluate the usefulness of our method's different components, we first perform an ablation study on the MMWHS dataset. We then compare our method against state-of-the-art UDA approaches on two segmentation tasks to show its superior performance.

4.5.1 Ablation study

Impact of entropy loss We first evaluate the effectiveness of the entropy loss by varying coefficient λ_{ent} in Eq. (4.1). Table 4.1 shows the cross-validation performance obtained with different values of λ_{ent} on the MMWHS dataset. $\lambda_{\text{ent}} = 0$ represents the case where we only use the supervised loss on source data and the joint distribution alignment loss. As can be seen, using ℓ_{ent} with a weight of $\lambda_{\text{ent}} = 1 \times 10^{-5}$ leads to a mean DSC improvement of 0.62%. Performance however degrades for higher weight values, as the model is then forced to become more confident even for incorrect predictions.

Table 4.1 Impact in terms of DSC (mean ± stdev) of the weight λ_{ent} of ℓ_{ent} on the output, when performing cross validation.

λ_{ent}	**LVM**	**LAC**	**LVC**	**AA**	**Mean**
Lower baseline	6.97 ± 2.52	31.85 ± 1.56	34.64 ± 13.00	52.57 ± 6.31	31.51
0	65.26 ± 3.62	82.86 ± 3.87	79.94 ± 4.06	82.70 ± 2.50	77.69
1×10^{-6}	65.75 ± 4.90	82.88 ± 2.50	82.10 ± 3.61	82.01 ± 1.76	78.18
1×10^{-5}	73.87 ± 4.88	82.03 ± 2.51	80.41 ± 4.38	76.91 ± 5.18	78.31
3×10^{-5}	69.21 ± 1.54	81.40 ± 0.79	81.10 ± 2.43	81.38 ± 1.04	78.27
5×10^{-5}	71.18 ± 2.92	83.09 ± 3.26	81.13 ± 3.09	75.30 ± 6.16	77.67
1×10^{-4}	66.88 ± 3.08	77.86 ± 1.72	82.27 ± 1.33	76.60 ± 1.77	75.90

Table 4.2 Impact in terms of DSC (mean ± stdev) of the weight λ of ℓ_{align} on the output, when performing cross validation.

λ_{align}	**LVM**	**LAC**	**LVC**	**AA**	**Mean**
0	39.08 ± 6.70	69.64 ± 3.36	60.41 ± 4.74	52.45 ± 8.76	55.39
1×10^{-5}	63.90 ± 2.82	81.01 ± 4.01	74.43 ± 4.35	77.36 ± 4.84	74.18
5×10^{-5}	68.59 ± 4.69	78.47 ± 5.30	76.99 ± 8.40	79.59 ± 2.86	75.91
1×10^{-4}	73.87 ± 4.88	82.03 ± 2.51	80.41 ± 4.38	76.91 ± 5.18	78.31
5×10^{-4}	69.07 ± 5.45	80.99 ± 2.51	79.65 ± 6.37	82.23 ± 4.67	77.98

Table 4.3 Impact in terms of DSC (mean ± stdev) of the displacement range for the output and Upconv2 layer.

Layer	**Displ.**	**DSC (%)**				
		LVM	**LAC**	**LVC**	**AA**	**Mean**
Output	Δ_0	69.38 ± 1.59	85.34 ± 2.59	82.46 ± 1.40	78.14 ± 5.33	78.83
	Δ_1	67.95 ± 2.00	84.91 ± 1.71	82.17 ± 0.99	80.48 ± 4.33	78.87
	Δ_3	69.22 ± 3.28	88.60 ± 1.84	81.81 ± 2.11	81.34 ± 2.79	80.24
	Δ_5	67.60 ± 1.71	87.29 ± 1.55	83.78 ± 2.45	82.22 ± 3.15	80.22
Upconv2	Δ_0	73.24 ± 3.05	88.72 ± 1.04	88.91 ± 1.02	79.68 ± 2.72	82.64
	Δ_1	75.26 ± 3.41	86.95 ± 1.49	87.38 ± 0.84	83.47 ± 1.95	83.26
	Δ_3	76.40 ± 4.68	87.72 ± 0.45	89.84 ± 0.41	84.63 ± 2.42	84.65
	Δ_5	73.95 ± 4.07	87.33 ± 0.58	87.49 ± 2.11	75.85 ± 6.53	81.16

Impact of alignment loss We evaluate the effectiveness of our shape-aware joint distribution alignment loss ℓ_{align} used only on segmentation outputs. For this experiment, we set $\lambda_{\text{align}}^{\text{cc}}$ and ℓ_{ent} in Eq. (4.1) to 0 and vary the value of coefficient λ_{align}. Table 4.2 reports the cross-validation performance obtained with different λ_{align} on the MMWHS dataset. We observe a steady rise in accuracy when increasing λ_{align} from 0 to 1×10^{-4}, reaching an improvement of 22.92% in mean DSC. This shows the effective guidance of our joint distribution alignment loss. Although a small drop in mean DSC is observed when further increasing the strength of alignment ($\lambda_{\text{align}}=5\times10^{-4}$), performance remains significantly higher than the case without alignment loss ($\lambda_{\text{align}}=0$).

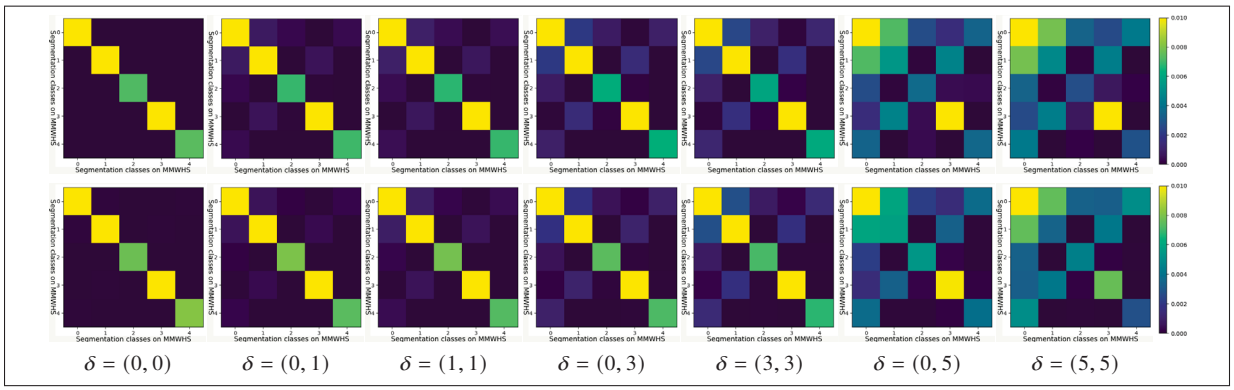


Figure 4.3 Joint matrix corresponding to different displacement vector δ, where $(0,0)$ corresponds to no displacement. The first row shows joint matrices from the source domain, and the second row joint matrices from the target domain.

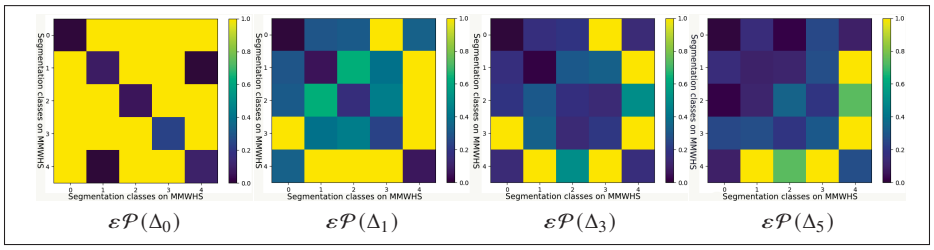


Figure 4.4 Cross-domain error of joint matrices, computed over displacement set Δ.

Impact of displacements Our method uses displacements to capture the shape of a class and the spatial relationships between different classes. We investigate the impact of using such

displacements in our joint distribution alignment and cross-correlation matrix alignment losses, respectively operating on the output and upconv2 layer. Table 4.3 reports the test performance with different displacement sets $\Delta_d = \{(w, h) \,|\, |w| \leq d \ \wedge \ |h| \leq d\}$, with displacement scale $d \in \{0, 1, 3, 5\}$. Both for the output space and upconv2, we achieve the highest mean DSC for displacement set Δ_3. Compared to having no displacement (Δ_0), this setting gives a mean DSC improvement of 1.41% for the output space and 2.01% for upconv2. As can be seen, using a larger displacements decreases performance, especially for the cross-correlation matrix alignment in upconv2 layer, which can be due to the greater uncertainty of relationships at larger distances.

Figure 4.3 shows examples of joint matrices for source and target domains, computed for different displacement vectors $\boldsymbol{\delta}$. As can be seen, different displacements give rise to distinct joint matrices, which provide unique information for alignment. We also observe that joint matrices are well-aligned across domains, for all displacements. To further measure alignment accuracy, we give in Figure 4.4 the relative difference between joint matrices for different displacement scales d, computed as follows: $\varepsilon\mathcal{P}(\Delta_d) = \sum_{\boldsymbol{\delta} \in \Delta_d} |\mathcal{P}^S(\boldsymbol{\delta}) - \mathcal{P}^T(\boldsymbol{\delta})| / \mathcal{P}^S(\boldsymbol{\delta})$. Without any displacement (Δ_0), differences mostly occur on off-diagonal elements, due to the fact that the relevant information (relative size of classes – i.e., class marginal) lies on the diagonal. On the other hand, for larger displacements, differences are also reduced for off-diagonal elements, which demonstrates the importance of spatial relationships between classes in this setting.

Impact of multi-scale computations Instead of using larger displacement scales, which would increase computations, our method proposes a multi-scale strategy where small displacement scales are used on output/feature maps down-sampled to different resolutions. To evaluate this strategy, we first compute the joint matrices from predictions down-sampled to a resolution of $\frac{1}{2}\times$, $\frac{1}{4}\times$ and $\frac{1}{8}\times$ the size of the original image. The *Output* portion of Table 4.4 reports the test performance of using multi-scale resolutions on the output space with displacement set Δ_1. As can be observed, the best performance is obtained when combining the first three resolutions (256, 128, and 64), and adding a lower resolution (32) decreases performance. This suggests that the joint matrix for low resolutions encodes relationships from unrelated regions, which

Table 4.4 Impact in terms of DSC (mean ± stdev) of the multi-resolution scales for the output and Upconv2 layer.

Layer	Multi-resolution				DSC (%)				
	256	128	64	32	LVM	LAC	LVC	AA	Mean
Output	✓				67.95 ± 2.00	84.91 ± 1.71	82.17 ± 0.99	80.48 ± 4.33	78.87
	✓	✓			68.18 ± 0.79	86.07 ± 1.59	81.53 ± 0.29	80.33 ± 1.02	79.03
	✓	✓	✓		70.51 ± 2.42	87.04 ± 0.76	82.86 ± 0.37	81.73 ± 3.45	80.53
	✓	✓	✓	✓	69.09 ± 2.79	87.10 ± 0.79	80.57 ± 2.73	82.81 ± 3.76	79.89
Upconv2	✓				75.26 ± 3.41	86.95 ± 1.49	87.38 ± 0.84	83.47 ± 1.95	83.26
	✓	✓			76.38 ± 2.10	87.79 ± 1.48	89.19 ± 0.74	85.20 ± 1.07	84.64
	✓	✓	✓		77.22 ± 4.76	87.73 ± 0.11	89.68 ± 1.19	84.63 ± 1.62	84.82
	✓	✓	✓	✓	77.15 ± 3.14	88.18 ± 0.40	89.46 ± 0.54	82.10 ± 3.11	84.22

Table 4.5 Impact of joint distribution matrix alignment and cross-correlation matrix alignment.

Output	Upconv2	LVM	LAC	LVC	AA	Mean
✗	✗	39.08 ± 6.70	69.64 ± 3.36	60.41 ± 4.74	52.45 ± 8.76	55.39
✓	✗	70.51 ± 2.42	87.04 ± 0.76	82.86 ± 0.37	81.73 ± 3.45	80.53
✗	✓	77.22 ± 4.76	87.73 ± 0.11	89.68 ± 1.19	84.63 ± 1.62	84.82
✓	✓	79.77 ± 1.72	89.47 ± 0.60	91.16 ± 0.82	87.02 ± 1.77	86.86

are not useful for segmentation. The same analysis was performed for the cross-correlation alignment on upconv2, see the *Upconv2* portion of Table 4.4.

Impact of using intermediate layers Next, we measure the impact of computing the alignment loss on an intermediate layer of the network (Upconv2 layer in Figure 4.2), in addition to the output layer. In a previous version of our method, we used a projection head to map the latent features to a discrete distribution over K clusters, and then computed the joint distribution alignment loss as done for the output space. However, this version led to a low performance as it lacks effective guidance for the unsupervised clustering. Figure 4.5 shows examples of clusters to be aligned across domains, for different values of K. We see that clusters corresponding to meaningful regions of the image are found for both domains, however there is no clear

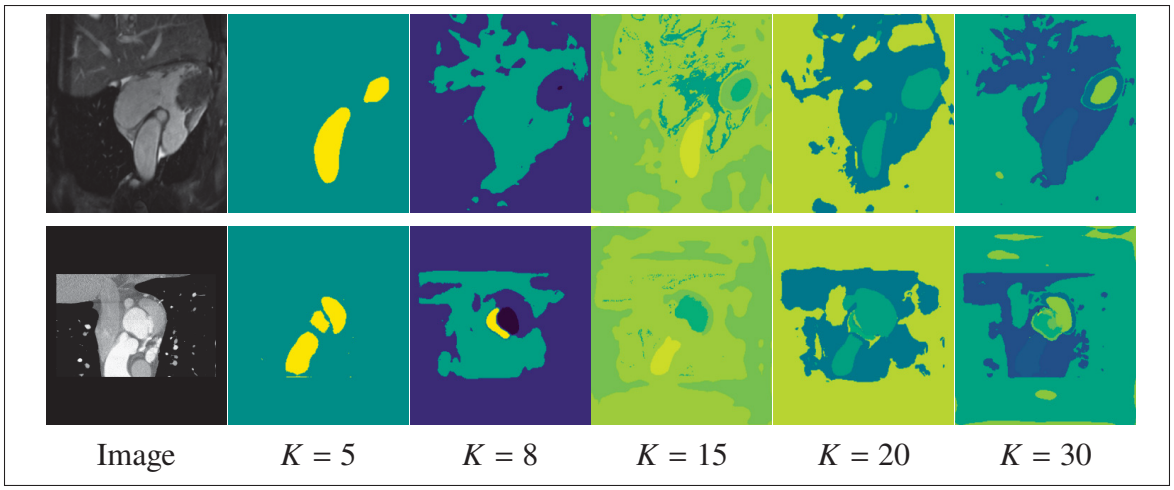


Figure 4.5 Clusters to be aligned across domains. The first row shows clusters from the source domain, and the second row clusters from the target domain.

correspondence of these clusters across domains. Moreover, due to the lack of supervised guidance, the clustering often collapses to unbalanced solutions where a few clusters dominate. In contrast, aligning features directly using our cross-correlation loss better preserves semantic information, as shown in Figure 4.6.

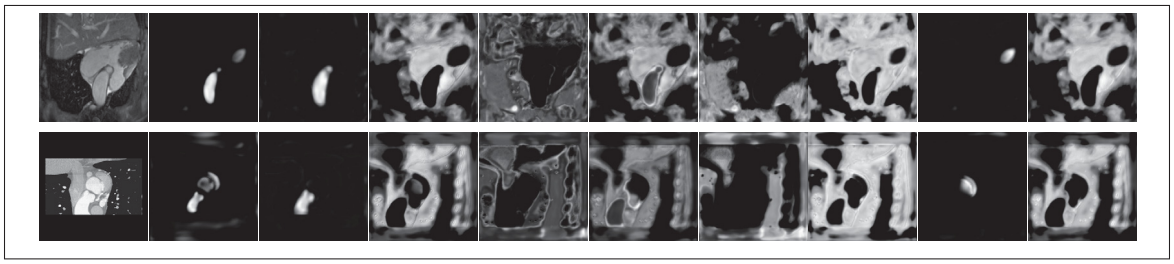


Figure 4.6 Features to be aligned across domain. The first row shows features from the source domain, and the second row features from the target domain. Columns 2-10 correspond to different feature maps.

To further illustrate the power of our alignment strategy on an intermediate layer, we show in Figure 4.7 a 2D t-SNE plot of feature vectors used to compute the cross-correlation matrix, as well as the alignment error (absolute difference) of cross-correlation matrices across domains. We see that the target data features obtained by EntDA are noisier and less structured than those learned by our method, especially for the LVM and AA classes. Using our cross-correlation

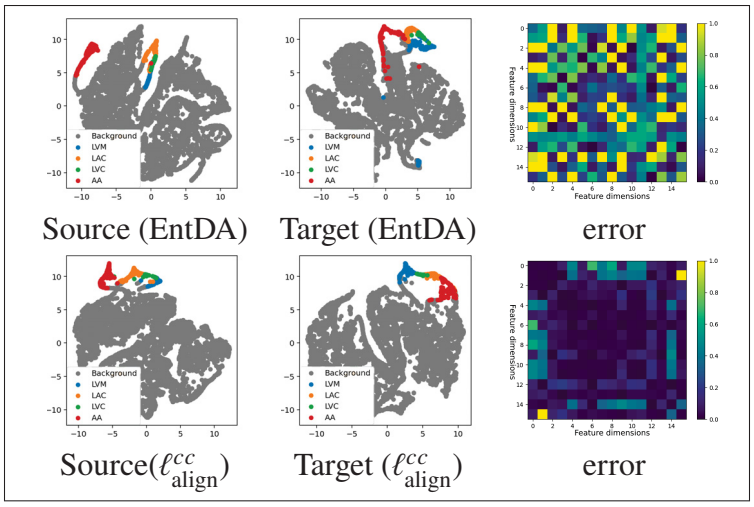


Figure 4.7 Comparison of t-SNE plot and alignment error (absolute difference) between EntDA and our cross-correlation alignment loss on the MMWHS dataset.

alignment loss $\ell^{\text{cc}}_{\text{align}}$, features of target domain examples are more clustered according to their real class. Our $\ell^{\text{cc}}_{\text{align}}$ loss also leads to more similar cross-correlation matrices across domains, demonstrating the better alignment of the intermediate representation.

In Table 4.5, we report the test performance obtained by performing alignment on the output layer, upconv2 layer, or both. Compared to having no alignment, the joint distribution alignment of the output improves mean DSC by 25.14%, and the cross-correlation alignment on upconv2 by 29.43%. However, a greater improvement 31.47% is achieved by combining the two alignment strategies, which shows their complementarity.

Results with different network architectures To evaluate the robustness of our method to different backbone architectures, we also employ ENet (Adam *et al.*, 2016) as underlying segmentation network. Table. 4.6 reports the DSC performance of our method, Ent + prior, and EntDA with different segmentation backbones. We see that our method gives similar improvements over Ent + prior for the two architectures, with a 4.38% DSC boost when using ENet and 6.71% boost for UNet. ENet was developed for the real-time semantic segmentation.

Table 4.6 Performance comparison of the proposed method with different domain adaptation methods for cardiac and prostate segmentation, in terms of DSC (mean ± stdev).

Architectures	Methods	MMWHS				
		LVM	LAC	LVC	AA	Mean
Enet	Upper baseline	88.46 ± 0.65	85.36 ± 6.14	91.39 ± 0.36	80.21 ± 2.40	86.36
	Lower baseline	6.04 ± 3.14	47.03 ± 19.88	25.11 ± 13.00	53.51 ± 15.74	32.92
	EntDA	12.42 ± 6.42	70.27 ± 7.12	54.64 ± 13.94	46.81 ± 8.47	46.04
	Ent+prior	61.41 ± 3.77	83.76 ± 2.38	71.36 ± 1.03	80.19 ± 2.98	74.18
	Ours	66.40 ± 1.99	81.83 ± 2.09	80.59 ± 0.17	85.22 ± 1.92	78.51
Unet	Upper baseline	89.07 ± 0.17	91.31 ± 0.18	92.33 ± 0.43	88.85 ± 2.52	90.39
	Lower baseline	6.97 ± 2.52	31.85 ± 1.56	34.64 ± 13.00	52.57 ± 6.31	31.51
	EntDA	39.08 ± 6.70	69.64 ± 3.36	60.41 ± 4.74	52.45 ± 8.76	55.39
	Ent+prior	73.72 ± 3.01	86.43 ± 1.69	80.42 ± 5.35	80.05 ± 1.72	80.15
	Ours	79.77 ± 1.72	89.47 ± 0.60	91.16 ± 0.82	87.02 ± 1.77	86.86

Table 4.7 Performance comparison of the proposed method with different domain adaptation methods for cardiac and prostate segmentation, in terms of DSC (mean ± stdev).

Methods	MMWHS					Prostate
	LVM	LAC	LVC	AA	Mean	
Upper baseline	89.07 ± 0.17	91.31 ± 0.18	92.33 ± 0.43	88.85 ± 2.52	90.39	88.11 ± 0.44
Lower baseline	6.97 ± 2.52	31.85 ± 1.56	34.64 ± 13.00	52.57 ± 6.31	31.51	68.68 ± 0.88
EntDA	39.08 ± 6.70	69.64 ± 3.36	60.41 ± 4.74	52.45 ± 8.76	55.39	69.61 ± 3.13
PLDA	38.88 ± 3.18	75.86 ± 1.05	66.34 ± 1.42	72.78 ± 2.90	63.47	72.27 ± 0.15
Ent+prior	73.72 ± 3.01	86.43 ± 1.69	80.42 ± 5.35	80.05 ± 1.72	80.15	71.28 ± 2.35
SIFA	59.79 ± 5.21	78.74 ± 1.88	74.38 ± 2.93	85.38 ± 1.78	74.57	73.94 ± 1.55
MT-UDA	53.84 ± 3.76	81.61 ± 0.78	71.07 ± 2.50	87.12 ± 2.29	73.41	70.44 ± 1.20
PointcloudDA	58.42 ± 2.33	81.30 ± 1.19	68.22 ± 4.01	84.13 ± 2.11	73.02	70.89 ± 1.24
Ours	79.77 ± 1.72	89.47 ± 0.60	91.16 ± 0.82	87.02 ± 1.77	86.86	74.97 ± 1.27

Although it is faster than UNet, it yields a lower segmentation accuracy, which is the reason why we chose UNet as our backbone architecture for the main results in Table 4.7.

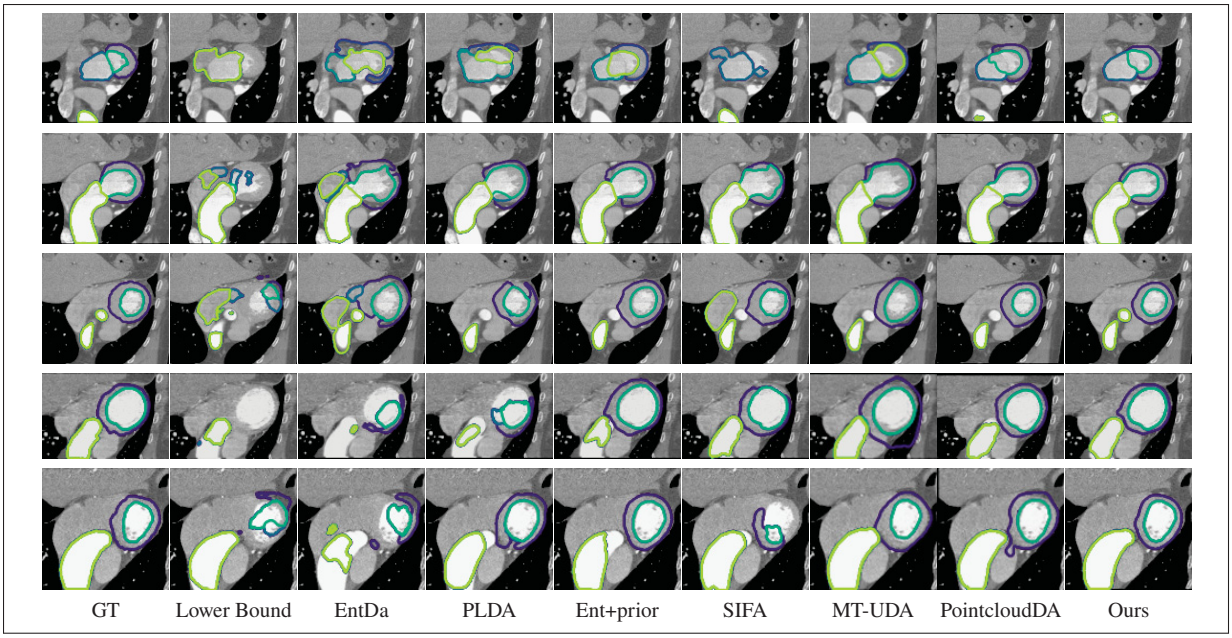


Figure 4.8 Visual comparison of methods with respect to the ground-truth (GT). Each row corresponds to a different CT image from the MMWHS test set. Purple: LVM; Blue: LAC; Dark green: LVC; Green: AA.

4.5.2 Comparison with the state-of-art

Table 4.7 compares our best performing variant with baselines and recent UDA approaches, on the cardiac and prostate segmentation tasks. For cardiac segmentation, the Lower-bound baseline which trains the network only on source (MR) images obtains a low DSC score of 31.51%. Among compared approaches, EntDA has a mean DSC improvement of 23.88%. PLDA, employing the well-trained model of EntDA to generate pseudo labels for target data, achieves a mean DSC of 63.47%, which is 8.08% higher than EntDA. SIFA, MT-UDA and PointcloudDA, which are three methods based on adversarial learning, achieve improvements of 43.06%, 41.90% and 41.51%, respectively, compared to the baseline. Ent + prior, aligning the class marginals (i.e., relative proportion of pixels in each class) of the source and target domains, obtains 49% improvements over the Lower-bound baseline. By aligning higher-order statistics encoding the shape of individual classes and the spatial relationship between classes, our method obtains the best DSC for all classes. Specifically, it yields a mean DSC of 6.71% higher than the

second best approach (Ent + prior), and a 55.35% improvement over the Lower-bound baseline. For the prostate segmentation task, our method also outperforms all other approaches, yielding a 6.29% higher mean DSC than the Lower-bound baseline and 1.03% higher than the second best method (SIFA).

Figure 4.8 shows some visual segmentation examples, comparing our proposed method with other approaches for MMWHS CT test images. While other approaches fail to properly segment the challenging LVM and LVC classes, our method makes predictions close to the ground-truth (GT) thanks to the alignment of high-order multi-scale statistics derived from the segmentation output and latent features.

4.6 Conclusion

We proposed a novel UDA method for cross-domain segmentation based on the alignment of high-order domain-invariant statistics encoding spatial relationships between classes. Our method estimates the joint class probability matrix for different spatial displacements and then aligns these matrices across the source and target domains. An efficient multi-scale strategy was proposed to capture long-range relationships with limited computations. We also propose to align the cross-correlation matrices for feature embeddings in intermediate layers of the network, which capture high-level semantic information. Extensive experiments were carried out on a multi-modal cardiac segmentation benchmark and prostate segmentation task. Experimental results confirmed the advantage of our approach over recently-proposed UDA methods for cross-domain image segmentation.

A potential limitation of our method is the need to compute joint and correlation matrices for the different domains, displacement vectors and scales, which increases training time. Fortunately, we found that optimal results could be obtained using a small set of displacements (Δ_1) in combination with the proposed multi-scale strategy. Moreover, the joint and cross-correlation matrices can be calculated efficiently in GPU using standard convolution operations, thereby limiting the computational overhead. Another possible drawback of our method is that it requires

a rough alignment of images from the source and target domains within a batch, during training. In the current implementation, this is achieved by selecting, for each batch, images that fall in the same randomly-chosen range of slices inside the corresponding volume. As future work, we will explore strategies to remove this dependency in the input space.

CONCLUSION AND RECOMMENDATIONS

This thesis focuses on medical image segmentation under two challenging problems: semi-supervised segmentation and unsupervised domain adaptation segmentation. After studying the body of work on these problems and discussing the limitations of existing approaches, we proposed methods to address these limitations, which represent three separate contributions.

Contribution 1: Self-paced and self-consistent co-training for semi-supervised image segmentation

Due to the large imbalance between labeled and unlabeled data in semi-supervised segmentation, predictions for unlabeled images are often imprecise, specially in the early stages of training. To tackle this issue and enhance performance, we first proposed a self-paced and self-consistent co-training method for segmentation that gradually adjusts the training process and ensures a consistent learning. Specifically, we developed a self-paced learning strategy for co-training that enables jointly-trained neural networks to focus on easier-to-segment regions first and then gradually consider harder ones. This is achieved via an end-to-end differentiable loss in the form of a generalized JSD. We also enhanced this generalized JSD loss with an uncertainty regularizer to encourage predictions from the co-trained networks to be both consistent and confident. Furthermore, to enhance the robustness of each co-trained model, we incorporated a self-consistency loss based on temporal ensembling that further strengthens the training by promoting consistent predictions across different model iterations. We evaluated the proposed method in three challenging segmentation tasks. Experimental results showed our method to outperform state-of-art approaches for semi-supervised segmentation and yield a performance close to full-supervision while using only a small fraction of the labeled data.

By employing self-paced and self-consistent strategies, the proposed method effectively addresses the issue of inaccurate predictions during the initial training stage, thereby significantly improving the performance and robustness of the semi-supervised segmentation network. However, multiple

segmentation networks are required to be constructed which increases the computational and memory requirements. Though using parallel computation techniques and employing a moving average update strategy can speed up training, the memory requirement of multiple models remains a concern. A potential solution to this problem is to create a co-trained framework that enables knowledge distillation within a single model across views or have the co-trained networks share some of their layers.

Contribution 2: CAT: constrained adversarial training for anatomically-plausible semi-supervised segmentation

In addition to a high accuracy, in our second contribution, we also consider the anatomically plausibility of the predicted segmentation as critical requirement. However, achieving this can be challenging in a semi-supervised setting due to the limited availability of labeled data. To alleviate this problem, we developed a constrained adversarial training method that avoids anatomically-invalid segmentations by integrating complex non-differentiable anatomical priors into the network. The proposed method incorporates constraints via an adversarial training strategy, which can generate constraint-violating examples for unlabeled images, and the REINFORCE algorithm employed to obtain useful gradients for non-differentiable constraints. Experiments on synthetic data and on four clinically-relevant datasets show the effectiveness of our method in terms of segmentation accuracy and anatomical plausibility.

The proposed method offers a generic and efficient way to add complex segmentation constraints on top of any segmentation network. While we have shown its usefulness for segmenting various organs, our method may not be suitable for segmenting regions of unknown shape such as lesions and tumors. Moreover, in this work, we explored well-known geometrical constraints such as shape connectivity and connectivity. As future work, one could explore a broader range of segmentation constraints related to anatomy.

Contribution 3: Shape-aware joint distribution alignment for cross-domain image segmentation

We proposed an UDA method for cross-domain segmentation, which achieves competitive performance by aligning high-order domain invariant statistics encoding spatial relationships between classes. Unlike existing approaches that rely on complex techniques like CycleGANs, our method explicitly aligns the domain-invariant representations via a regularization loss. As a result, it offers a simple yet highly effective solution for UDA. Our method first estimates the joint class probability matrix with different spatial displacements, which encode the domain-invariant information including shape size and spatial relationship between classes. Domain adaptation is then achieved by aligning these matrices across the source and target domains. Moreover, an efficient multi-scale strategy is introduced to capture long-range spatial relationships with low computation costs. We also proposed to align the cross-correlation matrices for feature embeddings in intermediate layers of the network, which captures the relationship between high-level semantic features. We tested our method on the task of unpaired multi-modal cardiac segmentation and prostate segmentation. Our results show the advantages of our method compared to recent approaches for cross-domain image segmentation.

The proposed method provides a simple and effective way to achieve domain adaptation. However, a possible drawback of this method is that it requires a rough alignment of input images from the source and target domains within a batch, during training. In the current implementation, this is achieved by selecting, for each batch, images that fall in the same randomly-chosen range of slices inside the corresponding volume. As future work, we could explore strategies to remove this dependency in the input space. Another limitation is that the domain-invariant representation learned by our method may not accommodate the segmentation of complex shapes like lesions. In subsequent research, one could extend our method beyond the segmentation of normal organs and tissues.

APPENDIX I

APPENDIX FOR PAPER «SELF-PACED AND SELF-CONSISTENT CO-TRAINING FOR SEMI-SUPERVISED IMAGE SEGMENTATION»

1. Proof of Theorem 1

Proof. With fixed $\{\theta^k\}$ and $\{\widehat{\mathbf{y}}_{ui}\}$, the optimal learning weights w_{uik} corresponding to pixel i of unlabeled example u and model f^k are found by solving

$$\min_{w_{uik}\in[0,1]} \frac{\gamma}{2} w_{uik}^2 + \left(D_{\mathrm{KL}}\left(\mathbf{p}_{ui}^k \,||\, \widehat{\mathbf{y}}_{ui}\right) - \gamma\right) w_{uik} \tag{A I-1}$$

If $D_{\mathrm{KL}}\left(\mathbf{p}_{ui}^k \,||\, \widehat{\mathbf{y}}_{ui}\right) \geq \gamma$, since $w_{uik} \geq 0$, the minimum is obviously achieved for $w_{uik}^* = 0$. Else, if $D_{\mathrm{KL}}\left(\mathbf{p}_{ui}^k \,||\, \widehat{\mathbf{y}}_{ui}\right) < \gamma$, we find the optimal weight by deriving the function w.r.t. w_{uik} and setting the result to zero, which gives

$$w_{uik}^* = 1 - \frac{1}{\gamma} D_{\mathrm{KL}}\left(\mathbf{p}_{ui}^k \,||\, \widehat{\mathbf{y}}_{ui}\right). \tag{A I-2}$$

Since both γ and $D_{\mathrm{KL}}\left(\mathbf{p}_{ui}^k \,||\, \widehat{\mathbf{y}}_{ui}\right)$ are non-negative, we have that $w_{uik}^* \in [0, 1]$, thus it is a valid solution. Considering both cases simultaneously, we therefore get

$$w_{uik}^* = \max\left(1 - \frac{1}{\gamma} D_{\mathrm{KL}}\left(\mathbf{p}_{ui}^k \,||\, \widehat{\mathbf{y}}_{ui}\right),\, 0\right) \tag{A I-3}$$

□

2. Proof of Theorem 2

Proof. Considering learning weights $\{\mathbf{w}_{ui}\}$ as fixed, optimizing model parameters $\{\theta^k\}$ and pseudo-labels $\{\widehat{y}_{ui}\}$ in (2.4) corresponds to:

$$\begin{aligned} &\min_{\{\theta^k\},\{\widehat{\mathbf{y}}_{ui}\}} \ \frac{1}{|\mathcal{U}|} \sum_{\mathbf{x}_u \in \mathcal{U}} \sum_{k=1}^{K} \sum_{i \in \Omega} w_{uik}\, D_{\mathrm{KL}}\big(\mathbf{p}_{ui}^k \,||\, \widehat{\mathbf{y}}_{ui}\big) \\ &\text{s.t.} \ \sum_{j \in C} \widehat{y}_{uij} \ = \ 1, \ \forall u, \forall i; \ \ \widehat{y}_{uij} \in [0,1], \ \forall u, \forall i, \forall j. \end{aligned} \tag{A I-4}$$

Since the pseudo-labels $\widehat{\mathbf{y}}_{ui}$ for each pixel are decoupled in the loss, we can optimize them independently. For each resulting sub-problem, we deal with the constraint that $\widehat{\mathbf{y}}_{ui}$ is a probability distribution with a Lagrangian formulation and convert the problem into

$$\begin{aligned} &\max_{\mu} \min_{\widehat{\mathbf{y}}_{ui}} \ \sum_{k=1}^{K} w_{uik}\, D_{\mathrm{KL}}\big(\mathbf{p}_{ui}^k \,||\, \widehat{\mathbf{y}}_{ui}\big) \ - \ \mu\Big(\sum_{j \in C} \widehat{y}_{uij} \ - \ 1\Big) \\ &= \ -\sum_{k=1}^{K} \sum_{j \in C} w_{uik}\, p_{uij}^k \log \frac{\widehat{y}_{uij}}{p_{uij}^k} \ - \ \mu\Big(\sum_{j \in C} \widehat{y}_{uij} \ - \ 1\Big) \end{aligned} \tag{A I-5}$$

where μ is the Lagrange multiplier corresponding to the one-sum constraint on $\widehat{\mathbf{y}}_{ui}$. Next, we derive this function with respect to each $\widehat{y}_{uij}$ and set the result to zero, yielding

$$\widehat{y}^{*}_{uij} \ = \ -\frac{1}{\mu} \sum_{k=1}^{K} w_{uik}\, p_{uij}^k \tag{A I-6}$$

To find μ we use the constraint that $\sum_j \widehat{y}_{uij} = 1$:

$$\begin{aligned} -\frac{1}{\mu} \sum_{j \in C} \sum_{k=1}^{K} w_{uik}\, p_{uij}^k \ &= \ 1 \\ \mu \ = \ -\sum_{j \in C} \sum_{k=1}^{K} w_{uik}\, p_{uij}^k \ &= \ -\sum_{k=1}^{K} w_{uik} \end{aligned} \tag{A I-7}$$

Using this equation in (A I-6) thus yields

$$\widehat{y}^{*}_{uij} \;=\; \frac{\sum_{k=1}^{K} w_{uik}\, p^{k}_{uij}}{\sum_{k=1}^{K} w_{uik}} \;=\; \sum_{k=1}^{K} \pi_{uik}\, p^{k}_{uij}. \tag{A I-8}$$

We finally insert (A I-8) in the loss of (A I-4) to get

$$\begin{aligned}
&\min_{\{\theta^{k}\},\{\widehat{\mathbf{y}}_{ui}\}} \; -\frac{1}{|\mathcal{U}|} \sum_{\mathbf{x}_u \in \mathcal{U}} \sum_{k=1}^{K} \sum_{j \in C} w_{uik}\, p^{k}_{uij}\, \log \frac{\widehat{y}_{uij}}{p^{k}_{uij}} \\
&= \; \frac{1}{|\mathcal{U}|} \sum_{\mathbf{x}_u \in \mathcal{U}} \sum_{i \in \Omega} \rho_{ui} \Big[-\sum_{j \in C} \big(\sum_{k=1}^{K} \pi_{uik}\, p^{k}_{uij} \big) \log \big(\sum_{k'=1}^{K} \pi_{uik'}\, p^{k}_{uij} \big) \\
&\quad + \; \sum_{k=1}^{K} \pi_{uik} \sum_{j \in C} p^{k}_{uij} \log\, p^{k}_{uij} \Big] \\
&= \; \frac{1}{|\mathcal{U}|} \sum_{\mathbf{x}_u \in \mathcal{U}} \sum_{i \in \Omega} \rho_{ui} \Big[\mathcal{H}\big(\sum_{k=1}^{K} \pi_{uik}\, \mathbf{p}^{k}_{ui} \big) \; - \; \sum_{k=1}^{K} \pi_{uik} \mathcal{H}\big(\mathbf{p}^{k}_{ui}\big) \Big] \\
&= \; \frac{1}{|\mathcal{U}|} \sum_{\mathbf{x}_u \in \mathcal{U}} \sum_{i \in \Omega} \rho_{ui}\, \mathrm{JSD}_{\boldsymbol{\pi}_{ui}}\big(\mathbf{p}^{1}_{ui}, \ldots, \mathbf{p}^{K}_{ui}\big).
\end{aligned} \tag{A I-9}$$

□

BIBLIOGRAPHY

Adam, P., Abhishek, C., Sangpil, K. & Eugenio, C. (2016). ENet: A Deep Neural Network Architecture for Real-Time Semantic Segmentation. *arXiv:*, abs/1606.02147.

Allah, A. M. G., Sarhan, A. M. & Elshennawy, N. M. (2023). Edge U-Net: Brain tumor segmentation using MRI based on deep U-Net model with boundary information. *Expert Systems with Applications*, 213, 118833.

Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Kopp-Schneider, A., Landman, B. A., Litjens, G., Menze, B., Ronneberger, O., Summers, R. M. et al. (2022). The medical segmentation decathlon. *Nature communications*, 13(1), 1–13.

Antti, T. & Valpola, H. (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Proceedings of the Neural Information Processing Systems*, pp. 1195-1204.

Arjovsky, M., Chintala, S. & Bottou, L. (2017). Wasserstein generative adversarial networks. *International conference on machine learning*, pp. 214–223.

Ba, J. L., Kiros, J. R. & Hinton, G. E. (2016). Layer normalization. *arXiv preprint arXiv:1607.06450*.

Badrinarayanan, V., Kendall, A. & Cipolla, R. (2017). SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12), 2481-2495.

Bai, W., Oktay, O., Sinclair, M., Suzuki, H., Rajchl, M., Tarroni, G., Glocker, B., King, A., Matthews, P. M. & Rueckert, D. (2017). Semi-supervised learning for network-based cardiac MR image segmentation. *Medical Image Computing and Computer-Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part II 20*, pp. 253–260.

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*, pp. 490–499.

Bateson, M., Dolz, J., Kervadec, H., Lombaert, H. & Ayed, I. B. (2021). Constrained domain adaptation for image segmentation. *IEEE Transactions on Medical Imaging*, 40(7), 1875–1887.

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

Bengio, Y., Louradour, J., Collobert, R. & Weston, J. (2009). Curriculum learning. *Proceedings of the International Conference on Machine Learning*, pp. 41-48.

Bernard, O., Alain, L., Clement, Z., Frederick, C., Xin, Y., Pheng, A. H., Irem, C. & *et al.* (2018). Deep Learning Techniques for Automatic MRI Cardiac Multi-structures Segmentation and Diagnosis: Is the Problem Solved? *IEEE Transaction on Medical Imaging*, 37(11), 2514-2525.

Blum, A. & Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. *Proceedings of ACM Computational Learning Theory*, pp. 92-100.

Bortsova, G., Dubost, F., Hogeweg, L., Katramados, I. & De Bruijne, M. (2019). Semi-supervised medical image segmentation via learning consistency under transformations. *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22*, pp. 810–818.

Boutillon, A., Conze, P.-H., Pons, C., Burdin, V. & Borotikar, B. (2021). Multi-task, multi-domain deep segmentation with shared representations and contrastive regularization for sparse pediatric datasets. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 239–249.

Chaitanya, K., Karani, N., Baumgartner, C. F., Becker, A., Donati, O. & Konukoglu, E. (2019). Semi-supervised and task-driven data augmentation. *Information Processing in Medical Imaging: 26th International Conference, IPMI 2019, Hong Kong, China, June 2–7, 2019, Proceedings 26*, pp. 29–41.

Chaitanya, K., Karani, N., Baumgartner, C. F., Erdil, E., Becker, A., Donati, O. & Konukoglu, E. (2021). Semi-supervised task-driven data augmentation for medical image segmentation. *Medical Image Analysis*, 68, 101934.

Chan, T. F. & Vese, L. A. (2001). Active contours without edges. *IEEE Transactions on image processing*, 10(2), 266–277.

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

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K. & Yuille, A. L. (2014). Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K. & Yuille, A. L. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4), 834–848.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F. & Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. *Proceedings of the European conference on computer vision (ECCV)*, pp. 801–818.

Cheplygina, V., de Bruijne, M. & Pluim, J. P. (2019). Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Medical Image Analysis*, 54, 280-296.

Comaniciu, D. & Meer, P. (2002). Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 24(5), 603–619.

Cui, W., Liu, Y., Li, Y., Guo, M., Li, Y., Li, X., Wang, T., Zeng, X. & Ye, C. (2019). Semi-supervised brain lesion segmentation with an adapted mean teacher model. *Information Processing in Medical Imaging: 26th International Conference, IPMI 2019, Hong Kong, China, June 2–7, 2019, Proceedings 26*, pp. 554–565.

Dai, P., Chen, P., Wu, Q., Hong, X., Ye, Q., Tian, Q., Lin, C.-W. & Ji, R. (2021a). Disentangling task-oriented representations for unsupervised domain adaptation. *IEEE Transactions on Image Processing*, 31, 1012–1026.

Dai, Y., Gao, Y. & Liu, F. (2021b). Transmed: Transformers advance multi-modal medical image classification. *Diagnostics*, 11(8), 1384.

Dasgupta, S., Littman, M. & McAllester, D. (2002). PAC generalization bounds for co-training. *Proceedings of the Neural Information Processing Systems*, pp. 375-382.

Dong, S., Luo, G., Tam, C., Wang, W., Wang, K., Cao, S., Chen, B., Zhang, H. & Li, S. (2020). Deep atlas network for efficient 3D left ventricle segmentation on echocardiography. *Medical image analysis*, 61, 101638.

Dou, Q., Liu, Q., Heng, P. A. & Glocker, B. (2020). Unpaired multi-modal segmentation via knowledge distillation. *IEEE Transaction on Medical Imaging*, 39(7), 2415-2425.

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

Duan, J., Bello, G., Schlemper, J., Bai, W., Dawes, T. J., Biffi, C., de Marvao, A., Doumoud, G., O'Regan, D. P. & Rueckert, D. (2019). Automatic 3D bi-ventricular segmentation of cardiac images by a shape-refined multi-task deep learning approach. *IEEE transactions on medical imaging*, 38(9), 2151–2164.

Dubuisson, M.-P. & Jain, A. (1994). A modified Hausdorff distance for object matching. *Proceedings of 12th International Conference on Pattern Recognition*, 1, 566-568 vol.1. doi: 10.1109/ICPR.1994.576361.

Englesson, E. & Azizpour, H. (2021). Generalized jensen-shannon divergence loss for learning with noisy labels. *Advances in Neural Information Processing Systems*, 34, 30284–30297.

Felzenszwalb, P. F. & Huttenlocher, D. P. (2004). Efficient graph-based image segmentation. *International journal of computer vision*, 59, 167–181.

Fralick, S. (1967). Learning to recognize patterns without a teacher. *IEEE Transactions on Information Theory*, 13(1), 57–64.

Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z. & Lu, H. (2019). Dual attention network for scene segmentation. *Proc. CVPR Conf.*, pp. 3146-3154.

Fu, M. C. (2006). Gradient estimation. *Handbooks in operations research and management science*, 13, 575–616.

Ganaye, P.-A., Sdika, M. & Benoit-Cattin, H. (2018). Semi-supervised learning for segmentation under semantic constraint. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 595–602.

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

Gao, Y., Huang, R., Yang, Y., Zhang, J., Shao, K., Tao, C., Chen, Y., Metaxas, D. N., Li, H. & Chen, M. (2021). FocusNetv2: Imbalanced large and small organ segmentation with adversarial shape constraint for head and neck CT images. *Medical Image Analysis*, 67, 101831.

Goetz, M., Weber, C., Binczyk, F., Polanska, J., Tarnawski, R., Bobek-Billewicz, B., Koethe, U., Kleesiek, J., Stieltjes, B. & Maier-Hein, K. H. (2015). DALSA: domain adaptation for supervised learning from sparsely annotated MR images. *IEEE transactions on medical imaging*, 35(1), 184–196.

Goodfellow, I., Bengio, Y. & Courville, A. (2016). *Deep learning*. MIT press.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. & Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, 63(11), 139–144.

Graham, R. L. & Yao, F. F. (1983). Finding the convex hull of a simple polygon. *Journal of Algorithms*, 4(4), 324–331.

Grandvalet, Y. & Bengio, Y. (2004). Semi-supervised learning by entropy minimization. *Advances in neural information processing systems*, 17.

Guan, H. & Liu, M. (2021). Domain adaptation for medical image analysis: a survey. *IEEE Transactions on Biomedical Engineering*.

Hang, W., Feng, W., Liang, S., Yu, L., Wang, Q., Choi, K.-S. & Qin, J. (2020). Local and global structure-aware entropy regularized mean teacher model for 3d left atrium segmentation. *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23*, pp. 562–571.

Hao, C., Xiaojuan, Q., Lequan, Y., Qi, D., Jing, Q. & Pheng-Ann, H. (2017). DCAN: Deep contour-aware networks for object instance segmentation from histology images. *Medical image analysis*, 36, 135–146.

Harold, J., Kushner, G. & Yin, G. (1997). Stochastic approximation and recursive algorithm and applications. *Application of Mathematics*, 35.

Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H. R. & Xu, D. (2022). Unetr: Transformers for 3d medical image segmentation. *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 574–584.

Hung, W.-C., Tsai, Y.-H., Liou, Y.-T., Lin, Y.-Y. & Yang, M.-H. (2018). Adversarial learning for semi-supervised semantic segmentation. *arXiv preprint arXiv:1802.07934*.

Ioffe, S. & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *International conference on machine learning*, pp. 448–456.

Jafari, M. H., Liao, Z., Girgis, H., Pesteie, M., Rohling, R., Gin, K., Tsang, T. & Abolmaesumi, P. (2019). Echocardiography segmentation by quality translation using anatomically constrained CycleGAN. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 655–663.

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

Jiang, L., Meng, D., Mitamura, T. & Hauptmann, A. G. (2014). Easy samples first: Self-paced reranking for zero-example multimedia search. *Proceedings of the 22nd ACM international conference on Multimedia*, pp. 547–556.

Kervadec, H., Dolz, J., Tang, M., Granger, E., Boykov, Y. & Ben Ayed, I. (2019). Constrained-CNN losses for weakly supervised segmentation. *Medical image analysis*, 54, 88–99.

Kingma, D. P. & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Krizhevsky, A., Sutskever, I. & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6), 84–90.

Kumagai, A. & Iwata, T. (2019). Unsupervised domain adaptation by matching distributions based on the maximum mean discrepancy via unilateral transformations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01), 4106–4113.

Kumar, M. P., Packer, B. & Koller, D. (2010). Self-paced learning for latent variable models. *Proceedings of the Neural Information Processing Systems*, pp. 1189-1197.

Laine, S. & Aila, T. (2016). Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*.

LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278–2324.

Lee, D.-H. et al. (2013). Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. *Workshop on challenges in representation learning, ICML*, 3(2), 896.

Li, C., Luo, X., Chen, W., He, Y., Wu, M. & Tan, Y. (2021). AttENT: Domain-Adaptive Medical Image Segmentation via Attention-Aware Translation and Adversarial Entropy Minimization. *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 952–959.

Lin, J. (1991). Divergence measures based on the Shannon entropy. *IEEE Transaction on Information theory*, 37(1), 145-151.

Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollár, P. (2017). Focal loss for dense object detection. *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988.

Litjens, G., Toth, R., van de Ven, W., Hoeks, C., Kerkstra, S., van Ginneken, B., Vincent, G., Guillard, G., Birbeck, N., Zhang, J. et al. (2014). Evaluation of prostate segmentation algorithms for MRI: the PROMISE12 challenge. *Medical Image Analysis*, 18(2), 359–373.

Liu, K., Tang, W., Zhou, F. & Qiu, G. (2019a). Spectral regularization for combating mode collapse in GANs. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6382–6390.

Liu, L., Jiang, H., He, P., Chen, W., Liu, X., Gao, J. & Han, J. (2019b). On the variance of the adaptive learning rate and beyond. *arXiv preprint arXiv:1908.03265*.

Long, J., Shelhamer, E. & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440.

Loshchilov, I. & Hutter, F. (2016). Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*.

Lu, J., Meng, D., Zhao, Q., Shan, S. & Hauptmann, A. (2015). Self-Paced Curriculum Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1).

Luc, P., Couprie, C., Chintala, S. & Verbeek, J. (2016). Semantic segmentation using adversarial networks. *arXiv preprint arXiv:1611.08408*.

Ma, F., Meng, D., Dong, X. & Yang, Y. (2020). Self-paced Multi-view Co-training. *Journal of Machine Learning Research*, 21(57), 1-38.

McCulloch, W. S. & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5, 115–133.

Miyato, T., Maeda, S., Koyama, M. & Ishii, S. (2019). Virtual Adversarial Training: A Regularization Method for Supervised and Semi-Supervised Learning. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 41(8), 1979-1993.

Mondal, A. K., Dolz, J. & Desrosiers, C. (2018). Few-shot 3d multi-modal medical image segmentation using generative adversarial learning. *arXiv:*, abs/1810.12241.

Mugnai, D., Pernici, F., Turchini, F. & Del Bimbo, A. (2022). Fine-grained adversarial semi-supervised learning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 18(1s), 1–19.

Nair, V. & Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. *Proceedings of the 27th international conference on machine learning (ICML-10)*, pp. 807–814.

Nosrati, M. S. & Hamarneh, G. (2016). Incorporating prior knowledge in medical image segmentation: A survey. *CoRR, arXiv:1607.01092*.

Oktay, O., Ferrante, E., Kamnitsas, K., Heinrich, M., Bai, W., Caballero, J., Cook, S. A., De Marvao, A., Dawes, T., O'Regan, D. P. et al. (2017). Anatomically constrained neural networks (ACNNs): application to cardiac image enhancement and segmentation. *IEEE transactions on medical imaging*, 37(2), 384–395.

Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N. Y., Kainz, B. et al. (2018). Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*.

Olsson, V., Tranheden, W., Pinto, J. & Svensson, L. (2021). Classmix: Segmentation-based data augmentation for semi-supervised learning. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1369–1378.

Ouyang, C., Kamnitsas, K., Biffi, C., Duan, J. & Rueckert, D. (2019). Data efficient unsupervised domain adaptation for cross-modality image segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 669–677.

Painchaud, N., Skandarani, Y., Judge, T., Bernard, O., Lalande, A. & Jodoin, P.-M. (2020). Cardiac segmentation with strong anatomical guarantees. *IEEE transactions on medical imaging*, 39(11), 3703–3713.

Pathak, D., Krahenbuhl, P. & Darrell, T. (2015). Constrained convolutional neural networks for weakly supervised segmentation. *Proceedings of the IEEE international conference on computer vision*, pp. 1796–1804.

Peng, J., Estrada, G., Pedersoli, M. & Desrosiers, C. (2020a). Deep co-training for semi-supervised image segmentation. *Pattern Recognition*, 107, 107269.

Peng, J., Kervadec, H., Dolz, J., Ayed, I. B., Pedersoli, M. & Desrosiers, C. (2020b). Discretely-constrained deep network for weakly supervised segmentation. *Neural Networks*, 130, 297–308.

Perone, C. S., Ballester, P., Barros, R. C. & Julien, C. (2019). Unsupervised domain adaptation for medical imaging segmentation with self-ensembling. *NeuroImage*, 194, 1-11.

Perone, C. S. & Cohen-Adad, J. (2018). Deep semi-supervised segmentation with weight-averaged consistency targets. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support* (pp. 12–19). Springer.

Qiao, S., Shen, W., Zhang, Z., Wang, B. & Yuille, A. (2018). Deep co-training for semi-supervised image recognition. *Proceedings of the European Conference on Computer Vision*, pp. 135-152.

Rasmus, A., Berglund, M., Honkala, M., Valpola, H. & Raiko, T. (2015). Semi-supervised learning with ladder networks. *Advances in neural information processing systems*, 28.

Reed, S., Lee, H., Anguelov, D., Szegedy, C., Erhan, D. & Rabinovich, A. (2014). Training deep neural networks on noisy labels with bootstrapping. *arXiv preprint arXiv:1412.6596*.

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*, pp. 234-241.

Rother, C., Kolmogorov, V. & Blake, A. (2004). " GrabCut" interactive foreground extraction using iterated graph cuts. *ACM transactions on graphics (TOG)*, 23(3), 309–314.

Rumelhart, D. E., Hinton, G. E. & Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, 323(6088), 533–536.

Scudder, H. (1965). Adaptive communication receivers. *IEEE Transactions on Information Theory*, 11(2), 167–174.

Shen, R., Tang, B., Lodi, A., Tramontani, A. & Ayed, I. B. (2020). An ILP model for multi-label MRFs with connectivity constraints. *IEEE Transactions on Image Processing*, 29, 6909–6917.

Shen, Z., Cao, P., Yang, H., Liu, X., Yang, J. & Zaiane, O. R. (2023). Co-training with High-Confidence Pseudo Labels for Semi-supervised Medical Image Segmentation. *arXiv preprint arXiv:2301.04465*.

Simonyan, K. & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Simpson, A. L., Michela, A. & *et al.* (2019). A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv:*, abs/1902.09063.

Souly, N., Spampinato, C. & Shah, M. (2017). Semi supervised semantic segmentation using generative adversarial network. *Proceedings of the International Conference on Computer Vision*, pp. 5689-5697.

Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S. & Jorge Cardoso, M. (2017). Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3*, pp. 240–248.

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

Tsai, Y.-H., Sohn, K., Schulter, S. & Chandraker, M. (2019). Domain adaptation for structured output via discriminative patch representations. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1456–1465.

Ulyanov, D., Vedaldi, A. & Lempitsky, V. (2016). Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*.

Vesal, S., Gu, M., Kosti, R., Maier, A. & Ravikumar, N. (2021). Adapt Everywhere: Unsupervised Adaptation of Point-Clouds and Entropy Minimization for Multi-Modal Cardiac Image Segmentation. *IEEE Transactions on Medical Imaging*, 40(7), 1838–1851.

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

Wachinger, C., Reuter, M., Initiative, A. D. N. et al. (2016). Domain adaptation for Alzheimer's disease diagnostics. *Neuroimage*, 139, 470–479.

Wang, W., Lu, Y., Wu, B., Chen, T., Chen, D. Z. & Wu, J. (2018). Deep active self-paced learning for accurate pulmonary nodule segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 723-731.

Wang, W., Li, H., Ding, Z. & Wang, Z. (2020). Rethink maximum mean discrepancy for domain adaptation. *arXiv preprint arXiv:2007.00689*.

Wang, Y., Zhang, Y., Liu, Y., Lin, Z., Tian, J., Zhong, C., Shi, Z., Fan, J. & He, Z. (2021a). ACN: adversarial co-training network for brain tumor segmentation with missing modalities. *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VII 24*, pp. 410–420.

Wang, Y., Peng, J. & Zhang, Z. (2021b). Uncertainty-aware pseudo label refinery for domain adaptive semantic segmentation. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9092–9101.

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4), 229–256.

Winn, J. & Shotton, J. (2006). The layout consistent random field for recognizing and segmenting partially occluded objects. *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, 1, 37–44.

Wu, F. & Zhuang, X. (2020). CF distance: A new domain discrepancy metric and application to explicit domain adaptation for cross-modality cardiac image segmentation. *IEEE Transactions on Medical Imaging*, 39(12), 4274–4285.

Wu, S., Chen, C., Xiong, Z., Chen, X. & Sun, X. (2021). Uncertainty-aware label rectification for domain adaptive mitochondria segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 191–200.

Wu, Y. & He, K. (2018). Group normalization. *Proceedings of the European conference on computer vision (ECCV)*, pp. 3–19.

Xia, Y., Liu, F., Dong, Y., Cai, J., Yu, L., Zhu, Z., Xu, D., Yuille, A. & Holger, R. (2020a). 3d semi-supervised learning with uncertainty-aware multi-view co-training. *Proceedings of the Applications of Computer Vision*, pp. 3646-3655.

Xia, Y., Yang, D., Yu, Z., Liu, F., Cai, J., Yu, L., Zhu, Z., Xu, D., Yuille, A. & Roth, H. (2020b). Uncertainty-aware multi-view co-training for semi-supervised medical image segmentation and domain adaptation. *Medical Image Analysis*, 65, 101766.

Xie, H., Fu, C., Zheng, X., Zheng, Y., Sham, C.-W. & Wang, X. (2023). Adversarial co-training for semantic segmentation over medical images. *Computers in Biology and Medicine*, 157, 106736.

Xie, Q., Luong, M.-T., Hovy, E. & Le, Q. V. (2020). Self-training with noisy student improves imagenet classification. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10687–10698.

Xu, C., Tao, D. & Xu, C. (2013). A survey on multi-view learning. *arXiv:*, abs/1304.5634.

Yang, J., Dvornek, N. C., Zhang, F., Chapiro, J., Lin, M. & Duncan, J. S. (2019a). Unsupervised domain adaptation via disentangled representations: Application to cross-modality liver segmentation. *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*, pp. 255–263.

Yang, J., Dvornek, N. C., Zhang, F., Zhuang, J., Chapiro, J., Lin, M. & Duncan, J. S. (2019b). Domain-agnostic learning with anatomy-consistent embedding for cross-modality liver segmentation. *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 0–0.

Yi-de, M., Qing, L. & Zhi-Bai, Q. (2004). Automated image segmentation using improved PCNN model based on cross-entropy. *Proceedings of 2004 International Symposium on Intelligent Multimedia, Video and Speech Processing, 2004.*, pp. 743–746.

Yu, L., Wang, S., Li, X., Fu, C.-W. & Heng, P.-A. (2019). Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation. *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*, pp. 605–613.

Zhang, J., Li, Z., Zhang, C. & Ma, H. (2020). Robust adversarial learning for semi-supervised semantic segmentation. *2020 IEEE International Conference on Image Processing (ICIP)*, pp. 728–732.

Zhang, Y., Yang, L., Chen, J., Fredericksen, M., Hughes, D. P. & Chen, D. Z. (2017). Deep adversarial networks for biomedical image segmentation utilizing unannotated images. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 408-416.

Zhang, Z., Yang, L. & Zheng, Y. (2018). Translating and segmenting multimodal medical volumes with cycle-and shape-consistency generative adversarial network. *Proceedings of the IEEE conference on computer vision and pattern Recognition*, pp. 9242–9251.

Zhao, A., Balakrishnan, G., Durand, F., Guttag, J. V. & Dalca, A. V. (2019). Data augmentation using learned transformations for one-shot medical image segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8543-8553.

Zhao, Q., Meng, D., Jiang, L., Xie, Q., Xu, Z. & Hauptmann, A. (2015). Self-paced learning for matrix factorization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1).

Zhao, Z., Xu, K., Li, S., Zeng, Z. & Guan, C. (2021). MT-UDA: Towards Unsupervised Cross-modality Medical Image Segmentation with Limited Source Labels. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 293–303.

Zheng, H., Lin, L., Hu, H., Zhang, Q., Chen, Q., Iwamoto, Y., Han, X., Chen, Y.-W., Tong, R. & Wu, J. (2019). Semi-supervised segmentation of liver using adversarial learning with deep atlas prior. *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22*, pp. 148–156.

Zheng, X., Fu, C., Xie, H., Chen, J., Wang, X. & Sham, C.-W. (2022). Uncertainty-aware deep co-training for semi-supervised medical image segmentation. *Computers in Biology and Medicine*, 149, 106051.

Zheng, Z., Wang, X., Zhang, X., Zhong, Y., Yao, X., Zhang, Y. & Wang, Y. (2020). Semi-supervised segmentation with self-training based on quality estimation and refinement. *Machine Learning in Medical Imaging: 11th International Workshop, MLMI 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Proceedings*, pp. 30–39.

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

Zhou, Y., Wang, Y., Tang, P., Song, B., Shen, W., Elliot, K. F. & Yuille, A. (2019b). Semi-supervised 3D abdominal multi-organ segmentation via deep multi-planar co-training. *Proceedings of the IEEE Workshop on Applications of Computer Vision*, pp. 121-140.

Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N. & Liang, J. (2018). Unet++: A nested u-net architecture for medical image segmentation. *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, pp. 3–11.

Zhu, J.-Y., Park, T., Isola, P. & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232.

Zhu, Z., He, X., Qi, G., Li, Y., Cong, B. & Liu, Y. (2023). Brain tumor segmentation based on the fusion of deep semantics and edge information in multimodal MRI. *Information Fusion*, 91, 376–387.

Zhuang, X., Li, L., Payer, C., Štern, D., Urschler, M., Heinrich, M. P., Oster, J., Wang, C., Smedby, Ö., Bian, C. et al. (2019). Evaluation of algorithms for multi-modality whole heart segmentation: an open-access grand challenge. *Medical image analysis*, 58, 101537.

Zotti, C., Luo, Z., Lalande, A. & Jodoin, P.-M. (2018). Convolutional neural network with shape prior applied to cardiac MRI segmentation. *IEEE journal of biomedical and health informatics*, 23(3), 1119–1128.

Zou, Y., Yu, Z., Kumar, V. & Wang, J. (2018). Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. *Proceedings of the European Conference on Computer Vision*, pp. 289-305.