Reduced Supervision Methods for Medical Image Segmentation

by

Jizong PENG

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

MONTREAL, JUNE 9, 2022

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 Automated Production, École de technologie supérieure

Mr. Mohamad Forouzanfar, President of the Board of Examiners
Department of Automated Production, École de technologie supérieure

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

Mr. Julien Cohen-Adad, External Examiner
Department of Electrical Engineering, Polytechnique Montreal

THIS THESIS WAS PRESENTED AND DEFENDED

IN THE PRESENCE OF A BOARD OF EXAMINERS AND THE PUBLIC

ON MAY 30, 2022

AT ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deep thankfulness to my two supervisors: Professors Christian Desrosiers and Marco Pedersoli. I thank them for their endless support, availability and deep insight throughout my studies. Owing to Prof. Desrosiers, I was able to transfer my field of study from fibre-optics telecommunications to computer vision and deep learning. This proved to be a tipping point in my life and opened new horizons to my professional career. Prof. Desrosiers helped me obtain three competitive scholarships without which it would have been hard for me to focus entirely on research. He also encouraged me to participate in multiple internships, both in Montreal and in China, where I could enhance my problem-solving, presentation and communication skills while working on complex real-life problems. Prof. Pedersoli guided my research by offering brilliant insights on a wide range of topics and helped me have access to sufficient computation resources for my work. I am particularly grateful for his kind assistance during his recovery from illness, by exchanging text messages and have video-conference calls about my on-going research.

In short, I have received tremendous support from their part during my research, helping me with the choice of research topics, motivation discussion, idea formulation, code implementation and debugging, result discussion, and paper writing. The many times when I discussed new ideas with my two professors, writing equations on a whiteboard before COVID-19 pandemic, is still fresh in my memory. Their suggestions and recommendations where key factors leading to the acceptation of my manuscripts. Through their help, I gradually realized that doing research in very dynamic areas such as deep learning and computer vision requires a strong mathematical background, competitive coding skills, high resilience to failure (as it is part of our daily life), and more importantly, creativity. I also feel grateful for their optimism in life and their serious attitude in science, which enabled me to grow as a researcher. A special thank to them for their strong support during the COVID-19 pandemic. During this difficult period, and under huge pressure, we kept our discussions online, came up with new ideas and finally made them to be published.

I would also like to thank members of the LIVIA Lab, who have helped me with configuring the server, improving ideas, writing papers, setting up experiments as well as organizing a journal club every two weeks. I particularly thank Hoel Kervadec, whose segmentation framework became one of my main references to learn Python, PyTorch and pattern design, as well as a guide on organizing experiments. Many of my works have been realized in collaboration with Professors Ismail Ben Ayed and Jose Dolz, who I would like to thank for their warm welcome, continuous support and lively discussion. Their rich insights have been of great help to accurately identify problems and find solutions. I also appreciate the great help given by Ping Wang on setting up experiments and writing papers. Thanks to all this support, I was able to make contributions in the field of medical image segmentation during my PhD.

In the spare time of my PhD studies, I have had the chance to contribute to the development of two companies in Montreal: QuindiTech and Haxio. In QuindiTech, I met brilliant colleagues and friends, Nemo Xiong, Yuchen Lu, Xuan Li, Pingao Wang and Peng Xu from Mila, McGill and Polytechnique Montreal. Together, we created a skin diagnostic application *Q-Skin* based on mobile cameras, which makes high-accuracy skin disease diagnosis available to a broader population and won the first place in North America at the Alibaba start-up competition 2018. I would like to express my appreciation to them for the opportunity and the confidence they have given me. On many occasions, we discussed ideas, best model design, coding practice, business plan and presentation strategy until the next day. I especially appreciated the moment where we promoted our product in front of reviewers and hundreds of curious customers, which made me feel more confident than ever.

In Haxio, I had the chance to develop computer vision prototypes from scratch for production line inspection using embedded cameras. I was fortunate to collaborate with mature software engineers and worked on multiple projects combining both research and engineering. For this precious experience of more than one year, I would like to express my sincere gratitude to Alexandre Gervais and Kevin Hillairet, who helped me to integrate seamlessly to the team, taught me how to collaborate with software engineers, and showed me the highest level of professionalism. Working together in this enjoyable atmosphere is a memory I will not forget.

In the last year of my PhD, I had the honor to work in the newly-created Shanghai AI Lab, which hosts many AI experts from China and around the world. For the first time, I witnessed how AI-based systems studied my research can improve both the accuracy and the efficiency of doctors in big hospitals. Countless patients benefited from our efficient segmentation algorithms, identifying potential lesions and tumors at very early stage from complex abdomen CT images. This valuable experience wouldn't have happened without the help of Dr. Yinan Chen and other colleagues. I was highly impressed by their efficiency and persistence, and am convinced they will one day lead the innovation of AI-empowered medical systems.

Again, I would like to thank the Fond de Recherche Scientique et Technologies du Québec (FRQNT) for offering me financial support throughout my studies. I also thank Compute Canada, its regional partner Calcul Québec, 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 would like to thank my parents. They gave birth to me at an old age due to difficult times and the one-child policy, and became retired when I was still in primary school. In the last ten years, I was able to obtain undergraduate and graduate degrees, and to become a doctoral candidate in another country. Unfortunately, I was not able to accompany them in this part of their life. This is one of my biggest regrets. The last part of the acknowledgement is thus dedicated to them: thank you for your support and sacrifice.

Méthodes de Supervision Réduites pour la Segmentation des Images Médicales

Jizong PENG

RÉSUMÉ

La segmentation d'images médicales est une étape de pré-traitement importante dans les systèmes de diagnostic assisté par ordinateur. Les méthodes basées sur les réseaux de neurones ont démontré des performances de pointe sur diverses tâches de segmentation avec différentes modalités d'image. Malgré leur succès sans précédent, les réseaux de neurones nécessitent généralement une grande quantité de données étant étiquetées avec précision. Cependant, obtenir ces données est un processus laborieux et coûteux qui nécessite souvent l'intervention d'un expert médical, et les annotations sont sujettes aux erreurs. Pour mitiger la rareté des images densément annotées, une direction prometteuse de recherche consiste à exploiter des images avec des signaux de supervision réduits. Ces supervisions réduites se composent généralement d'une étiquette d'image, des points, des traits ou des boîtes englobantes comme annotation, cependant des images sans aucune information supervisée peuvent également être employées. De plus, des recherches récentes ont également tenté de combiner ces annotations faibles avec des *a priori* anatomiques de régions d'intérêt pour guider la prédiction du réseau vers des solutions anatomiquement plausibles.

L'objectif principal de cette thèse est de développer des algorithmes précis pour la segmentation d'images médicales, pouvant apprendre avec une supervision réduite. Plus précisément, nous proposons d'abord un algorithme de segmentation faiblement supervisé, apprenant à partir de traits et de contraintes anatomiques discrètes. Ensuite, nous présentons une approche de segmentation basée sur l'apprentissage par ensemble, permettant l'entraînement collaboratif de plusieurs réseaux de segmentation avec un nombre limité d'images étiquetées et une plus grande quantité d'images non étiquetées. Dans une autre contribution de la thèse, nous résolvons ce problème en introduisant un algorithme basé sur l'information mutuelle, qui emploi des images non étiquetées pour régulariser la représentation apprise par le réseau et augmente la précision de la segmentation lorsque peu d'images sont densément annotées. Par la suite, nous proposons une méthode basée sur l'apprentissage de la représentation qui exploite l'information d'images médicales non annotées avec des méta-étiquettes. Enfin, nous démontrons une méthode de maximization de l'information sensible aux contours pour le pre-entraînement des représentations denses du réseau, pouvant exploiter l'information sur les structures anatomiques d'images non étiquettées and ainsi améliorier de manière significative la précision de segmentation étant donné un petit ensemble d'images annotées. Cette thèse a donné lieu à trois articles de revues, deux articles dans des conférences avec comité de lecture, deux articles dans des séminaires en imagerie médicale, ainsi qu'à un article en cours d'évaluation. Les objectifs spécifiques de cette thèse sont présentés ci-dessous.

Comme *premier* objectif, nous proposons une stratégie efficace de segmentation faiblement supervisée pour imposer des contraintes ou des *a priori* de régularisation sur les régions cibles. Cette méthode de segmentation est une des premières à employer une optimisation discrète

avec un réseau de neurones, ce qui lui permet d'obtenir une solution plus rapidement et avec une plus grande précision. La méthode proposée repose sur l'algorithme de la méthode des multiplicateurs à direction alternée (ADMM) et entraîne un CNN avec des contraintes discrètes et des *a priori* de régularisation. La performance de cette méthode est validée sur la segmentation d'images médicales n'ayant que quelques pixels annotés, ainsi que des contraintes discrètes sur la taille et la régularité des frontières de régions à segmenter. Des expériences sur deux jeux de données de référence démontrent que notre méthode apporte des améliorations significatives par rapport aux approches existantes en termes de précision de segmentation, de satisfaction des contraintes et de vitesse de convergence.

Dans notre *deuxième* objectif, nous nous concentrons sur la segmentation semi-supervisée et proposons un algorithme basé sur l'apprentissage par ensemble. Cet algorithme entraîne plusieurs modèles avec un nombre réduit d'images annotées, ainsi que des images non annotées servant à échanger des informations entre les modèles. Afin d'assurer la diversité des modèles, une fonction de perte antagoniste est conçue. L'efficacité de notre méthode est démontrée sur trois tâches de segmentation d'images médicales couvrant différentes modalités, où celle-ci augmente la précision de segmentation lorsque très peu d'images étiquetées sont utilisées. L'effet de notre perte de diversité est également étudié en visualisant les images générées lors de l'entraînement antagnoiste. Nous explorons aussi le gain de performance obtenu avec un ensemble ayant plus de deux modèles, montrant que l'ajout de modèles améliore les résultats au coût de calculs accrus.

Dans notre *troisième* objectif, une nouvelle méthode de segmentation semi-supervisée est proposée. Cette méthode tire parti de l'information mutuelle sur les distributions catégorielles pour obtenir à la fois une invariance de représentation globale et une régularité spatiale de la segmentation. Dans cette méthode, nous maximisons l'information mutuelle pour les caractéristiques intermédiaires qui sont extraites à la fois de l'encodeur et du décodeur d'un réseau de segmentation. Une perte sur l'information mutuelle globale est employée sur l'encodeur pour favoriser l'invariance par rapport à des transformations géométriques sur les images d'entrée. De même, une perte sur l'information mutuelle locale est proposée pour encourager la cohérence spatiale dans les cartes de caractéristiques du décodeur, et ainsi fournir une segmentation plus régulière. Les avantages de notre méthode sont évalués sur quatre bases de données publiques pour la segmentation d'images médicales. Les résultats expérimentaux montrent que notre méthode surpasse les approches récentes de segmentation semi-supervisée, et fournit une précision proche de celle obtenue avec une supervision complète, tout en nécessitant très peu d'images annotées.

Dans notre *quatrième* objectif, nous visons à obtenir une représentation utile à partir d'images non étiquetées. Plus précisément, nous adaptons l'apprentissage contrastif pour entraîner l'encodeur du réseau dans différentes tâches prédéfinies: déterminer si deux images d'un volume IRM proviennent de la même position, de la même personne, ou si celles-ci ont été acquises au même instant du cycle cardiaque. Afin d'atténuer le bruit présent dans ces méta-étiquettes, une stratégie efficace d'apprentissage auto-rythmée est ensuite proposée pour l'apprentissage contrastif, ce qui se traduit par une représentation plus robuste et donc des améliorations en

performance pour les tâches de segmentation. Nous vérifions la qualité de la méthode proposée sur cinq jeux de données portant sur la segmentation d'images médicales, indiquant clairement l'avantage de notre mécanisme d'apprentissage auto-rythmé utilisant les méta-étiquettes.

Enfin, le *dernier* objectif spécifique de cette thèse présente une méthode basée sur le partitionnement de données pour apprendre une représentation discriminative pour les cartes de caractéristiques denses du réseau. Cette approche utilise une perte d'information mutuelle améliorée pour regrouper les caractéristiques denses en plusieurs partitions équilibrées et confiantes. Une perte sensible aux contours, basée sur la corrélaion croisée au niveau de pixels, est également utilisée pour aligner les régions de haute entropie du partitionnement avec les arêtes dans l'image, ce qui force les différentes partitions à correspondre aux structures anatomiques présentes dans l'image. Nos pertes proposées complémentent la perte contrastive supervisée présentée dans l'objectif précédent, et leur combinaison conduit à d'improtantes améliorations de performance pour la segmentation. Les résultats expérimentaux obtenus à partir de deux jeux de données cliniquement pertinents indiquent clairement l'avantage de notre méthode par rapport aux approches existantes à base d'apprentissage contrastif, conduisant à une précision de segmentation proche de celle de la supervision complète, mais avec seulement quelques exemples densément annotés.

Mots-clés: segmentation d'images médicales, segmentation semi-supervisée, segmentation faiblement supervisée, supervision réduite, apprentissage de la représentation

Reduced Supervision Methods for Medical Image Segmentation

Jizong PENG

ABSTRACT

Medical image segmentation is an important pre-processing step in computer-aided diagnosis systems. Methods based on neural networks have demonstrated state-of-the-art performance on various segmentation tasks with different image modalities. Despite their unprecedented success, neural networks usually require a large amount of reliable densely-labeled data. However, obtaining this data is a laborious and costly process, which often requires medical experts, and annotations obtained by this process can be prone to errors. To mitigate the scarcity of densely-annotated data, a promising research direction is to exploit images with reduced supervision signals. These reduced types of supervision usually consist of image tags, points, scribbles or bounding boxes as annotations, however images without any form of supervision can also be leveraged. Recent works have also tried to combine these weak annotations with anatomical priors for regions of interest to guide the network prediction towards anatomically-plausible solutions.

The main objective of this thesis is to develop accurate algorithms for medical image segmentation which can learn with reduced supervision. Specifically, we first propose a weakly-supervised segmentation algorithm that learns from scribbles and discrete anatomical constraints. Next, we present a segmentation framework, based on deep ensemble learning, that enables the collaborative training of multiple segmentation networks with a small set of labeled images and a larger amount of unlabeled ones. In another contribution of the thesis, we solve this problem by introducing an algorithm based on mutual information that uses unlabeled images to regularize the feature representation in the network and boost segmentation accuracy when few images are densely annotated. We then propose a method based on representation learning that exploits the information from unlabeled images with various medical meta-labels. As the last contribution, we demonstrate a boundary-aware information maximization method for dense representation pre-training, which acquires meaningful anatomical structure cues from unlabeled images and thus significantly improving segmentation accuracy given a small set of labeled images. This thesis has resulted in three journal publications, two papers in peer-reviewed international conferences, two short papers presented in medical imaging workshops, as well as one paper currently under review. The specific objectives of this thesis are presented below.

As our *first* objective, we propose an efficient strategy for weakly-supervised segmentation to impose constraints or regularization priors on target regions. This segmentation method is among the first to employ discrete optimization with a neural network, which enables the network obtain a more accurate solution faster. Our proposed method is based on the alternating direction method of multipliers (ADMM) algorithm and trains a CNN with discrete constraints and regularization priors. The performance of this method is validated on the segmentation of medical images with few annotated pixels, as well as discrete constraints of the size and boundary regularity of segmented regions. Experiments on two benchmark datasets showed

our method to provide significant improvements compared to existing approaches in terms of segmentation accuracy, constraint satisfaction and convergence speed.

In our *second* objective, we focus on semi-supervised segmentation and propose an algorithm based on ensemble learning. This method trains multiple models with a reduced number of annotated images, as well as with non-annotated images used for exchanging information between the trained models. To enforce the diversity of models, an adversarial loss is also designed. The effectiveness of this method is assessed on three medical image segmentation tasks covering different modalities, where it boosts segmentation accuracy when very few labeled images are used. The impact of our diversity loss is studied by visualizing the images generated by the adversarial training. We also explore the performance gains obtained with an ensemble containing more than two models, showing that adding models can improve results at the cost of increased computations.

In our *third* objective, a novel semi-supervised segmentation method is proposed. This method leverages the mutual information computed on categorical distributions to achieve both global representation invariance and spatial smoothness. In this method, we maximize the mutual information for intermediate feature embeddings that are taken from both the encoder and decoder of a segmentation network. A loss on global mutual information is employed on the encoder to enforce invariance towards geometric transformations. Likewise, a loss on the local mutual information is also used to promote spatial consistency in feature maps from the decoder, and thus to provide a smoother segmentation. The advantages of our method are evaluated on four challenging publicly-available datasets for medical image segmentation. Experimental results show our method to outperform recently-proposed approaches for semi-supervised segmentation and provide an accuracy near to full supervision while requiring very few annotated images.

In our *fourth* objective, we aim to acquire a useful representation by employing unlabeled images. Specifically, we adapt standard contrastive learning to train the encoder of the network for different pre-defined tasks: determining if two images of a 3D MRI scan are from the same position, same subject, or were acquired at the same moment of the cardiac cycle. In order to mitigate the noise presented in these meta-labels, an effective self-paced learning strategy is then proposed in contrastive learning, which yields a more robust representation and thus performance improvements for the segmentation tasks. We verify the quality of the proposed method on five medical image segmentation datasets, indicating clearly the advantage of our proposed self-paced mechanism using the meta-labels.

We present, in our *last* objective, a cluster-based method to learn discriminative representations for dense feature maps. This approach employs an improved mutual information loss to group dense embeddings into multiple balanced and confident clusters. A boundary-aware loss based on pixel-wise cross-correlation is also enforced to align the cluster boundaries to image edges, which regularizes different clusters to correspond to anatomical structures in the image. Our proposed losses complement the contrastive loss presented in the previous objective, and their combination leads to remarkable improvements for the downstream segmentation tasks. Experimental results obtained from two clinically-relevant benchmark datasets clearly indicate

the advantage of our method over contrastive-based counterparts, leading to a segmentation precision close to that of full-supervision, given only a few densely-annotated examples.

Keywords: medical image segmentation, semi-supervised segmentation, weakly-supervised segmentation, reduced supervision, representation learning

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

LIST OF ALGORITHMS

LIST OF ABBREVIATIONS

CNN	Convolutional Neural Networks
FCN	Fully Convolutional Network
SGD	Stochastic Gradient Descent
CT	Computed Tomography
MRI	Magnetic Resonance Imaging
CE	Cross Entropy
DSC	Sørensen–Dice Coefficient
HD	Hausdorff Distance
MIL	Multiple instance learning
CAM	Class Activation Map
MRF	Markov Random Fields
CRF	Conditional Random Field
DAP	Deep Atlas Prior
SDM	Signed Distance Map
AE	Auto-Encoder
VAE	Variational Auto-Encoder
MI	Mutual Information
GAN	Generative Adversarial Network
KLD	Kullback–Leibler Divergence
JSD	Jensen-Shannon Divergence
ADMM	Alternating Direction Method of Multipliers
CT	computed tomography
MRI	magnetic resonance imaging
DSC	Sørensen–Dice coefficient

LIST OF SYMBOLS AND UNITS OF MEASUREMENTS

θ	Network's parameter
$\mathcal{D}_{\text{train}}$	Training Dataset
$\mathcal{D}_{\text{val}}$	Validation Dataset
$\mathcal{D}_{\text{test}}$	Test Dataset
$\mathcal{L}(\cdot,\cdot)$	Loss Function
$\mathcal{L}_{\text{DSC}}$	DSC Loss
$\mathcal{L}_{\text{sup}}(\cdot,\cdot)$	Supervised Loss
$\mathcal{L}_{\text{spv}}(\cdot,\cdot)$	Supervised Loss
$\mathcal{L}_{\text{cot}}(\cdot,\cdot)$	Ensemble Agreement Loss
$\mathcal{L}_{\text{cons}}(\cdot,\cdot)$	Consistency Loss
$\mathcal{L}_{\text{div}}(\cdot,\cdot)$	Diversity Loss
$\hat{\mathbf{y}}$	Discrete CRF-regularized Proposal
$\tilde{\mathbf{y}}$	Discrete Size-regularized Proposal
L_2	L_2 Loss (MSE Loss)
$D_{\text{KL}}(\cdot,\cdot)$	KL-divergence
r_{adv}	Adversarial Noise
$\mathcal{H}(\cdot)$	Entropy
$\mathcal{H}(\cdot,\cdot)$	Cross Entropy
$\mathcal{L}_{\text{MI}}^{\text{global}}$	Global MI Loss
$\mathcal{L}_{\text{MI}}^{\text{local}}$	Local MI Loss
$\mathcal{L}_{\text{unsupCon}}$	Unsupervised Contrastive Loss

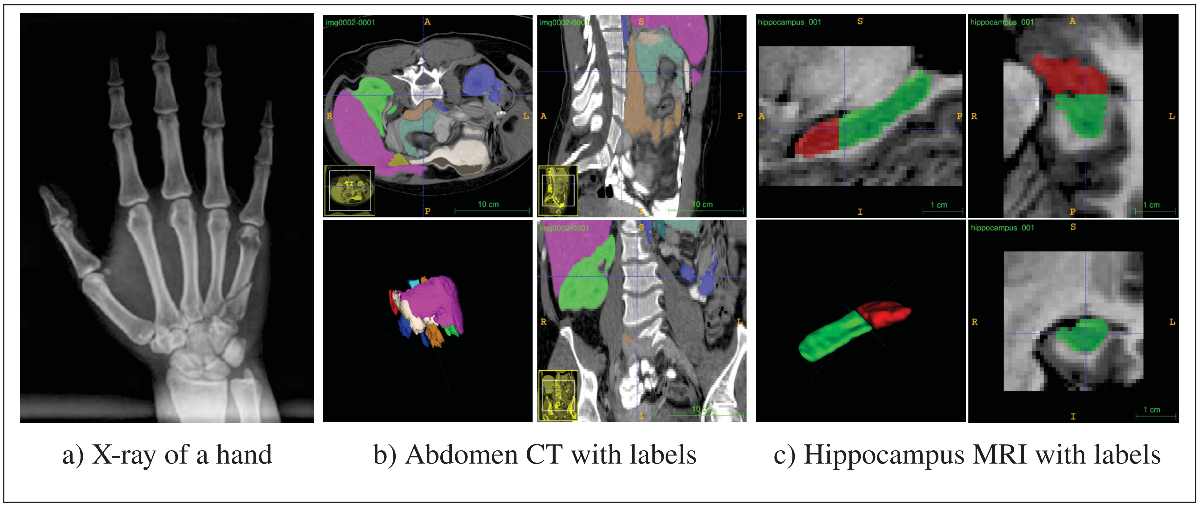
INTRODUCTION

0.1 Context

Medical imaging has become an essential tool for the non-invasive visualization of the body and plays a key role in the diagnosis, staging and monitoring of various diseases (Suetens, 2017; Zhang, Smith & Webb, 2008). Depending on the acquisition process and modalities, numerous types of medical images are available for clinical purpose, including X-ray (Röntgen, 1896), computed tomography (CT) (Hounsfield, 1980), magnetic resonance imaging (MRI) (Lauterbur, 1973), and so on. As illustrated in Fig. 0.1, these images provide invaluable information to clinicians about the different organs and tissues in the body. Given the complexity of such images, separating and identifying the regions of interest in these images, a problem known as *semantic segmentation*, is a critical step to evaluate the progression of a disease and plan future treatments (Pham, Xu & Prince, 2000).

0.2 Semantic segmentation for medical images

Semantic segmentation (Long, Shelhamer & Darrell, 2015) is an important task towards image understanding, which paves the way to automatic medical and satellite image analysis, and autonomous vehicles. Unlike image recognition which seeks to identify objects that appear in an image and whose output can be one or several discrete choices, semantic segmentation performs a *dense* classification by assigning a discrete label to *every pixel* based on their semantic meanings. Therefore, image segmentation considers both the class of an object as well as its precise location in the image, which requires to combine together contextual and spatial information. Semantic segmentation has been widely utilized in various applications, such as organ separation, tumour and lesion identification, as well as surgical planning, which play an essential role in medical image analysis.



a) X-ray of a hand b) Abdomen CT with labels c) Hippocampus MRI with labels

Figure 0.1 Visualization of different medical image modalities. a) X-ray of a hand, taken from (Al-Ayyoub *et al.*, 2013). b) multi-organ segmentation for an abdomen CT scan, image taken from the MICCAI 2015 Multi-Atlas Abdomen Labeling Challenge (Landman *et al.*, 2015). c) Hippocampus MRI segmentation, image taken from the Medical Segmentation Decathlon (Antonelli *et al.*, 2021). Both CT and MRI scans are visualized by ITK-SNAP (http://www.itksnap.org).

Among the various approaches employed for medical image segmentation, those based on deep learning have obtained the most success. These models, largely-known as neural networks, differ significantly from conventional approaches by adopting a *hierarchical architecture* of processing layers and learning representative features directly from *massive data*. At its core, a neural network discovers the hierarchical features behind big data by finding a high-dimensional and non-linear mapping from the input (e.g., raw image) to the output (e.g., its ground-truth label). The knowledge learned from data, under the form of parameters (or weights) of the neural network, is obtained via an iterative algorithm by minimizing a loss objective that forces the network's output to be as close as possible to the ground truth (Rumelhart, Hinton & Williams, 1986).

0.3 Challenges and problem statement

Despite their undeniable success in various medical imaging tasks, neural networks usually require a large amount of densely annotated (i.e., annotated at the *pixel* level) ground truth images. However, the manual or semi-automated annotation of data can be a very expensive step in the development of an intelligent system, especially for image segmentation. While class labels can be acquired within seconds for image classification, segmentation requires labeling each pixel of an input image which may take hours even for an experienced annotator. As example, annotating a single image of the well-known Cityscapes dataset takes about 1.5 hours in average (Cordts *et al.*, 2016). The annotation process is even more costly and challenging for medical image segmentation, since it requires to delineate complex regions of interest, such as organs and tumours, often in three dimensional space. Moreover, annotations embed the knowledge of experienced clinical experts, hence must made by well-trained doctors or radiologists. Further, high quality annotations should also be verified by multiple annotators, further reducing the availability of these experts for clinical diagnosis and research.

Another challenge arising for medical imaging tasks is that training datasets should contain a sufficient variability, for example different acquisition protocols and parameters across vendors, which can be hard to collect by a single institute. Sharing medical data across institutes and/or hospitals can be one solution but this raises legal and ethical issues (Ng, Lan, Yao, Chan & Feng, 2021). Therefore, the scarcity of densely annotated and varied data becomes a major constraint for the wide application of deep learning algorithms in medical imaging. Another significant challenge of medical image segmentation is the class imbalance problem (Yeung, Sala, Schönlieb & Rundo, 2021): in some scenarios, the foreground regions that we seek to identify (e.g., small tumours or specific bio-markers) can be orders of magnitude less frequent than the background pixels, resulting in a weak neural network that can hardly balance its precision and recall.

In light of these challenges, this thesis focuses on **the development of segmentation methods for medical images that reduces the burden of annotating images and are robust to variability and class imbalance in such images**.

0.4 Reduced supervision in medical image segmentation

Medical image segmentation techniques requiring less supervision have attracted a lot of attention in recent years. Learning from reduced supervision means that a neural network can learn to segment regions of interest with weak labels that are faster and cheaper to obtain, or by using images without any labels.

Weak annotations can have diverse forms based on the context of the target tasks. *Image-level*

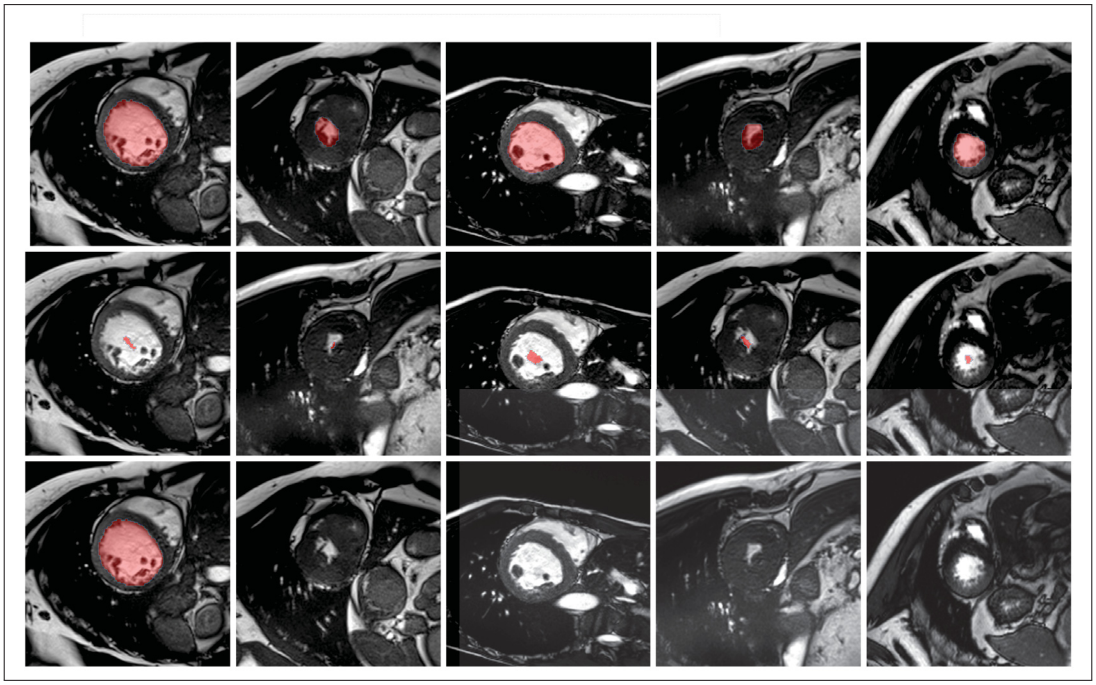


Figure 0.2 ACDC dataset with reduced annotation. Top row: Fully dense annotation of the left ventricle (LV) class; Middle row: Partial (Weak) annotation in the form of scribbles. Red-colored pixels indicate the presence of LV class at the corresponding locations; Bottom row: Annotation in a semi-supervised setting which combines a few fully-labeled images with a larger amount of unlabeled images.

tags is one of the simplest types of annotation for medical images and often comes for free by extracting diagnostic information from the clinical report using automatic algorithms. *Points* or *scribbles* are also widely used types of weak annotation, which require a radiologist to delineate a few pixels within the region of interests (Can *et al.*, 2018). Fig. 0.2 (top and middle rows) shows examples of dense annotations and scribbles in a cardiac segmentation dataset (Duane *et al.*, 2017). *Bounding boxes*, defined as the tightest box including the target region (Kervadec, Dolz, Wang, Granger & Ayed, 2020) are also commonly used for medical images. The learning process exploiting these weak labels is often known as "weakly-supervised learning".

It is apparent from Fig. 0.2 that weak annotations do not offer detailed spatial information for the target regions to segment. To mitigate this limitation, recent studies have also focused on incorporating medical knowledge, known as priors, as anatomical constraints to regularize the segmentation predictions. Prior knowledge can include information concerning the shape (Mirikharaji & Hamarneh, 2018), size range (Kervadec *et al.*, 2018) or topology (Clough *et al.*, 2019) of regions. Many of these constraints are defined in discrete space and can be solved by efficient discrete optimization. However, the discrete nature of these constraints represents a challenge for standard neural networks (Siu, Chan & Ming Lui, 2020; Nguyen *et al.*, 2018) which require computing the gradient of a continuous function.

Another way to leverage reduced supervision is to combine a small number of labeled data with a larger amount of label-free data, a scenario known as "semi-supervised learning". Unlike in weakly-supervised learning, no labels are provided for this additional data, and the network can only exploit the intrinsic characteristics of this data to improve both its segmentation accuracy and generalization capability. Semi-supervised segmentation methods are popular for medical image analysis because no additional annotation effort is required and any performance gain conferred by using unlabeled data comes with a low cost. However, the major challenge of this learning scenario lies in how to efficiently and thoroughly exploit these unlabeled data. Most works on this topic focus on designing effective regularization losses on unlabeled data and optimize this loss jointly with the supervised loss defined on labeled data. Various semi-supervised approaches exploit this idea during training, including consistency-based (Bortsova, Dubost, Hogeweg, Katramados & de Bruijne, 2019; Perone & Cohen-Adad, 2018), adversarial-based (Kim, Tack & Hwang, 2020), co-training based (Zhou *et al.*, 2018a) and representation-based (Chaitanya, Erdil, Karani & Konukoglu, 2020) methods.

0.5 Motivations and objectives

As highlighted previously, the main objective of this research thesis is developing novel segmentation algorithms to reduce the data annotation requirements for an accurate medical image segmentation system. We tackle this main objective by elaborating on five specific objectives. The first objective aims to leverage weak annotation and prior knowledge in a discrete framework, leading to better segmentation performance and strict constraint satisfaction. The second objective seeks to improve the segmentation quality in a semi-supervised setting by employing multiple segmentation models in an interactive way. This interactive learning can help individual models to learn from each other. The third objective, which also focuses on semi-supervised segmentation, designs a new regularization loss on different intermediate feature maps to boost the segmentation accuracy given limited labeled data. Our fourth objective instead exploits representation learning to define a self-paced representation learning scheme which can automatically learn from inaccurate meta information, and leads to a better generalization performance in a semi-supervised segmentation scenario. Our last objective further explores the concept of representation learning by exploiting pixel-wise clustering for dense representation pre-training. Our proposed method, which employs mutual information maximization and boundary-preserving learning, offers a highly-interpretable pre-trained model and pushes the segmentation accuracy close to full supervision, when fine-tuning the model on a few pixel-wisely labeled images. These objectives can be detailed as following:

Objective 1: Our *first* objective is to present an efficient learning framework which combines weakly-supervised annotations with discrete anatomical constraints. Many anatomical priors in medical image segmentation, such as the size range of a region and the length of its boundary, can be expressed as discrete constraints. However, the discrete nature of these constraints impedes their integration in a neural network. Previous works enforced the relaxation on these constraints, which often led to weak constraint satisfaction and sub-optimal solution. We aim to propose a novel method for training a CNN with discrete constraints and regularization priors. This method can be applied to the segmentation of medical images with scribbles as weak annotations, where both size constraints and boundary length regularization are enforced. We expect this method to significantly improve the segmentation accuracy and constraint satisfaction in a weakly-supervised setting.

Objective 2: Our *second* objective explores the combination of deep models to improve segmentation performance in semi-supervised setting with limited labeled data. While a single segmentation model often over-fits a small training data and predicts inaccurate segmentation boundaries for unseen ones, multiple models are expected to correct their predictions in an interactive manner. Our goal is to design losses encouraging the agreement and a diversity of models in the ensemble, which also favors the exchange of knowledge among the different models. Moreover, we seek to demonstrate the advantage of our method experimentally and compare it against state-of-the-art approaches for the semi-supervised segmentation of medical images.

Objective 3: The *third* objective of this thesis is to define a new regularization loss function based on mutual information to improve segmentation performance given limited labeled data. Mutual information is an information metric that measures the dependency relationship between two random variables. For this objective, we investigate the usefulness of mutual information measured on unlabeled images to achieve both global representation invariance and local feature regularity for semi-supervised image segmentation. Toward this goal, we will consider the intermediate feature maps from both the encoder and the decoder of the network given unlabeled images, as these features are important to final predictions and have been largely ignored so far. We explore two losses based on mutual information: a global loss constraining the encoder to learn an image representation that is invariant to geometric transformations and a local loss to promote spatial consistency in the feature maps of the decoder. We expect this method to outperform recently-proposed approaches for semi-supervised segmentation while training with very few annotated images.

Objective 4: Our *fourth* objective explores a representation learning framework employing two-stage training strategy: *pre-train* and *fine-tune*. In a first *pre-train* stage, a network is trained with a pre-defined (*pretext*) task that is different from the downstream segmentation task and does not require dense annotations, to acquire a useful representation. The pre-trained network is then fine-tuned in a second stage using a small set of labeled data. Previous works (Chaitanya *et al.*, 2020; Zeng *et al.*, 2021) in this direction defined the pre-training task as a classification problem using meta-labels on 2D images (slices) in volumetric images, for example predicting if two images are from the same patient or are located in a similar position inside their corresponding volume. However, because these meta-labels can be noisy, pre-training the network on this task can lead to a poor representation. Our fourth objective aims to mitigate this issue with a self-paced learning strategy that focuses the representation learning on confident meta-labels in the initial steps of training.

Objective 5: Our *last* objective, which also focuses on representation learning, is to develop a more effective way for pre-training the dense feature maps in the *decoder* of a segmentation network. Methods employing contrastive learning need to define pairs of positive and negative examples in a pre-training task. However, this can be challenging for dense feature maps without having access to dense annotations (Chaitanya *et al.*, 2020; Hu, Zeng, Xu & Shi, 2021). To overcome this problem, we will investigate an unsupervised representation learning approach leveraging a boundary-aware information maximization loss on dense feature maps. This approach seeks to group dense feature vectors into balanced and confident clusters that better align with anatomical regions in the image. It uses mutual information maximization to avoids trivial solutions and leverages a boundary-aware loss based on the cross-correlation to align the spatial entropy of clusters with edges in the image. The clusters obtained in the proposed unsupervised pre-training strategy are expected to be more representative of actual anatomic structures in the image, and therefore to boost accuracy for the downstream segmentation task given a few labeled examples.

The research related to these objectives has led to *three* published articles in journals (*Pattern Recognition*, *Neural Networks*, and *Machine Learning for Biomedical Imaging*), *two* papers presented in international conferences (*NeuRIPS* and *MIDL*), *two* short articles presented in medical imaging workshops, as well as one paper currently under review in a top machine learning venue (*ICML*).

0.6 Thesis outline

We split the thesis into a background chapter and several contribution chapters. In the following, we briefly introduce the content of each chapter.

Chapter 1 – Background

We start with a background chapter that introduces useful concepts and notation needed to understand the context and challenges of the problems addressed in the thesis. Background concepts include the basics of neural networks for medical image segmentation, as well as the datasets and metrics used to measure the performance of the proposed algorithms. We also detail the different types of labels and priors used in weakly-supervised image segmentation methods.

Then, we present related works on medical image segmentation with reduced supervision, including prior-based solutions exploiting weak labels and regularization-based techniques using unlabeled data.

Chapter 2 – Discretely-constrained deep network for weakly supervised segmentation

An efficient strategy for weakly-supervised segmentation is to impose constraints or regularization priors on target regions. Recent efforts have focused on incorporating such constraints in the training of convolutional neural networks (CNN), however this has so far been done within a continuous optimization framework. Yet, various segmentation constraints and regularization priors can be modeled and optimized more efficiently using a discrete formulation. In this chapter, we propose a method, based on the alternating direction method of multipliers (ADMM) algorithm, to train a CNN with discrete constraints and regularization priors. This method is applied to the segmentation of medical images with weak annotations where both size constraints and boundary length regularization are enforced. Experiments on two benchmark datasets for medical image segmentation show our method to provide significant improvements compared to existing approaches in terms of segmentation accuracy, constraint satisfaction and convergence speed.

The work in this chapter was initially presented at the Medical Imaging Meets NeurIPS Workshop in 2018, and then extended into a paper published in the Neural Networks journal in 2019.

Chapter 3 – Deep co-training for semi-supervised image segmentation

In this chapter, we aim to improve the performance of semantic image segmentation in a semi-supervised setting where training is performed with a small amount of annotated images and additional non-annotated images. We present a method based on an ensemble of deep segmentation models, where models are trained on subsets of the annotated data and use non-annotated images to exchange information with each other. Diversity across models is enforced with the use of adversarial samples. We demonstrate the potential of our method on three challenging image segmentation problems, and illustrate its ability to share information between simultaneously trained models, while preserving their diversity. Results indicate clear advantages in terms of performance compared to recently proposed semi-supervised methods for segmentation.

The content of this chapter was published in the Pattern Recognition journal in 2020.

Chapter 4 – Boosting Semi-supervised Image Segmentation with Global and Local Mutual Information Regularization

In this chapter, we present a semi-supervised segmentation method that leverages mutual information (MI) on categorical distributions to achieve both global representation invariance and local smoothness. The proposed method maximizes the MI for intermediate feature embeddings that are taken from both the encoder and decoder of a segmentation network. We first propose a global MI loss constraining the encoder to learn an image representation that is invariant to geometric transformations. Instead of resorting to computationally-expensive techniques for estimating the MI on continuous feature embeddings, we use projection heads to map them to a discrete cluster assignment where MI can be computed efficiently. Our method also includes a local MI loss to promote spatial consistency in the feature maps of the decoder and to provide a smoother segmentation. Since mutual information does not require a strict ordering of clusters in two different assignments, we incorporate a final consistency regularization loss on the output which helps align the cluster labels throughout the network. We evaluate the method on four challenging publicly-available datasets for medical image segmentation. Experimental results show our method to outperform recently-proposed approaches for semi-supervised segmentation and to yield an accuracy near to full supervision while training with very few annotated images.

The contributions related to this chapter resulted in two peer-reviewed papers, one presented at the 2020 Medical Imaging and Deep Learning (MIDL) conference and the other published in a special issue of the Machine Learning for Biomedical Imaging (MELBA) journal.

Chapter 5 – Self-Paced Contrastive Learning for Semi-supervised Medical Image Segmentation with Meta-labels

The contrastive pre-training of a neural network on a large dataset of unlabeled data often boosts the network's performance on downstream tasks like image classification. However, in domains such as medical imaging, collecting unlabeled data can be challenging and expensive. In the next chapter, we consider the task of medical image segmentation and adapt contrastive learning with meta-label annotations to scenarios where no additional unlabeled data is available. Meta-labels, such as the location of a 2D slice in a 3D MRI scan, often come for free during the acquisition process. We use these meta-labels to pre-train the image encoder, as well as in a semi-supervised learning step that leverages a reduced set of annotated data. A self-paced learning strategy exploiting the weak annotations is proposed to further help the learning process and discriminate useful labels from noise. Results on five medical image segmentation datasets show that our approach: *i*) highly boosts the performance of a model trained on a few scans, *ii*) outperforms previous contrastive and semi-supervised approaches, and *iii*) reaches close to the performance of a model trained on the full data.

The content of this chapter was published as regular paper in the 2021 Conference on Neural Information Processing Systems (NeurIPS).

Chapter 6 – Boundary-aware Information Maximization for Self-supervised Medical Image Segmentation

Unsupervised pre-training was shown to be an effective approach to boost various downstream tasks given limited labeled data. Contrastive learning is well-known method based on this idea that learns a discriminative representation by constructing positive and negative pairs. However, building such pairs for a segmentation task in an unsupervised way is not trivial. In this work, we propose a novel unsupervised pre-training framework that avoids the drawback of contrastive learning. Our framework builds on two principles: i) unsupervised image segmentation based on mutual information maximization as a pre-train-task, ii) and boundary-aware representation learning. Experimental results on two benchmark medical segmentation datasets reveal our method's effectiveness in improving segmentation performance when few annotated images are available.

The work presented in this chapter was submitted to the International Conference on Machine Learning (ICML) 2022.

Appendix I – Information-based Deep Clustering: An Experimental Study

Recently, two methods have shown outstanding performance for clustering images and jointly learning the feature representation. The first, called Information Maximizing Self-Augmented Training (IMSAT), maximizes the mutual information between input and clusters while using a regularization term based on virtual adversarial examples. The second, named Invariant Information Clustering (IIC), maximizes the mutual information between the clustering of a sample and its geometrically transformed version. These methods use mutual information in distinct ways and leverage different kinds of transformations. This Appendix presents a comprehensive analysis of transformation and losses for deep clustering, where we compare different combinations of these two components and evaluate how they interact with one another.

Results suggest that mutual information between a sample and its transformed representation leads to state-of-the-art performance for deep clustering, especially when used jointly with geometrical and adversarial transformations.

This analysis and its results are the subject of pre-print uploaded to Arvix in 2020.

Appendix II – Diversified Multi-prototype Representation for Semi-supervised Segmentation

In this second Appendix, we consider semi-supervised segmentation as a dense prediction problem based on prototype vector correlation and propose a simple way to represent each segmentation class with multiple prototypes. To avoid degenerate solutions, two regularization strategies are applied on unlabeled images. The first one leverages mutual information maximization to ensure that all prototype vectors are considered by the network. The second explicitly enforces prototypes to be orthogonal by minimizing their cosine distance. Experimental results on two benchmark medical segmentation datasets demonstrate our method's ability to improve segmentation accuracy using few annotated images.

The context of this preliminary work was presented at 2021 the Medical Imaging Meets NeurIPS Workshop.

Appendix III – Supplementary Material for Self-paced Contrastive Learning for Semi-supervised Medical Image Segmentation with Meta-labels

This supplementary material offers detailed explanation of the experimental setup, dataset description and complementary experimental results for Chapter 5 – "Self-Paced Contrastive Learning for Semi-supervised Medical Image Segmentation with Meta-labels", which was accepted at the NeurIPS 2021 conference.

Appendix IV – Appendix for Boundary-aware Information Maximization for Self-supervised Medical Image Segmentation

This appendix provides a conceptual diagram of our proposed method, detailed experimental setup, dataset description, data pre-processing steps, and complementary experimental results for Chapter 6 – "Boundary-aware Information Maximization for Self-supervised Medical Image Segmentation", which was submitted to the ICML 2022 conference.

Appendix V – Code Availability

We published most of our research code to support reproducibility and to promote the open-source spirit in the computer vision and AI communities. This appendix presents the code link for selected projects.

0.7 Publication summary

My PhD work has contributed to a total of 9 peer-reviewed papers (as well as one pre-print and one under review), 8 for which I am first author:

Manuscripts under review

- J. Peng, P. Wang, C. Desrosiers, M. Pedersoli, "Boundary-aware Information Maximization for Self-supervised Medical Image Segmentation", submitted to International Conference on Machine Learning (ICML) 2022. **(#3 conference in artificial intelligence, h5-index:204)**

Journal papers

- J. Peng, H. Kervadec, J. Dolz, I. Ben Ayed, M. Pedersoli, C. Desrosiers, “Discretely-constrained deep network for weakly supervised segmentation”, Neural Networks, volume 130, 2019. **(impact factor: 8.05)**

- J. Peng, G. Estrada, M. Pedersoli, C. Desrosiers, “Deep co-training for semi-supervised image segmentation”, Pattern Recognition, volume 107, 2020. **(impact factor: 7.74)**

- J. Peng, M. Pedersoli, C. Desrosiers, “Boosting Semi-supervised Image Segmentation with Global and Local Mutual Information Regularization”, Machine Learning for Biomedical Imaging, volume 1, 2021. **(selected for the MIDL 2020 special issue)**

- P. Wang, J. Peng, M. Pedersoli, ,Y. Zhou, C. Zhang, C. Desrosiers, “Self-paced and self-consistent co-training for semi-supervised image segmentation”, Medical Image Analysis, volume 73, 2021. **(impact factor: 8.54)**

Conference papers

- J. Peng, P. Wang, C. Desrosiers, and M. Pedersoli, “Self-Paced Contrastive Learning for Semi-supervised Medical Image Segmentation with Meta-labels”, Conference on Neural Information Processing Systems (NeurIPS) 2021. **(#1 conference in artificial intelligence, h5-index: 245)**

- J. Peng, M. Pedersoli, C. Desrosiers, “Mutual information deep regularization for semi-supervised segmentation”, Medical Imaging with Deep Learning (MIDL), PMLR, 2020.

- P. Wang, J. Peng, M. Pedersoli, Y. Zhou, C. Zhang, C. Desrosiers, “Context-aware virtual adversarial training for anatomically-plausible segmentation”, International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2021.

Workshop papers

- J. Peng, C. Desrosiers, M. Pedersoli, "DC-SegNet: A discretely constrained deep network for weakly supervised segmentation", Medical Imaging meets NeurIPS, 2018.

- J. Peng, C. Desrosiers, M. Pedersoli, "Diversified Multi-prototype Representation for Semi-supervised Segmentation", Medical Imaging meets NeurIPS, 2021.

Preprints

- J. Peng, C. Desrosiers, M. Pedersoli, "Information based Deep Clustering: An experimental study", arXiv, 1910.01665, 2019.

CHAPTER 1

BACKGROUND

In this chapter, we cover the basic neural network components and architectures that are used in medical image segmentation tasks, as well as the different loss functions employed to optimize these architectures. We then present related works proposed to reduce the requirement of dense annotation for accurate medical image segmentation, including those using weak annotation and prior knowledge, as well as methods using label-free data.

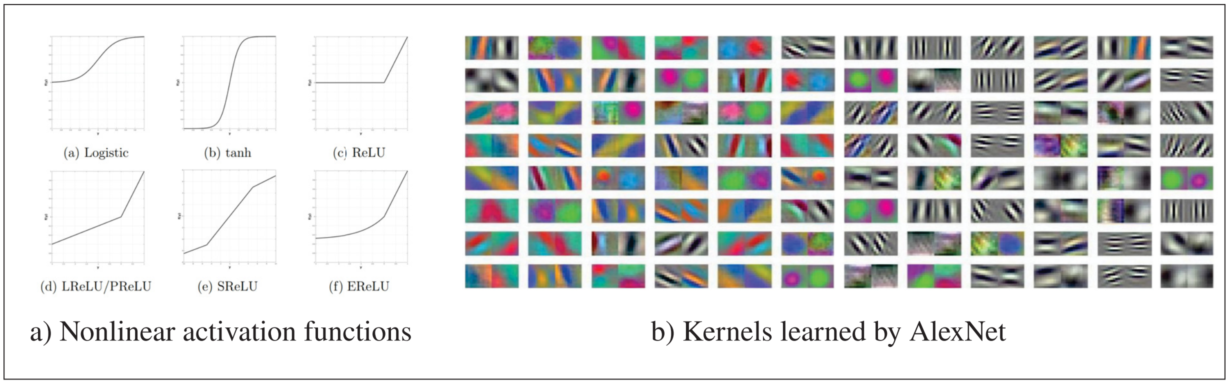
1.1 Basic blocks for neural networks

Neural networks offer a powerful solution to learn complex relationships between given inputs and target outputs for various tasks (LeCun, Bengio & Hinton, 2015). By adding more layers and more units within a layer, a deep network can represent functions of increasing complexity. Modern neural networks can be composed of several main blocks, each of which has its own design purpose. In what follows, We briefly introduce the blocks most commonly used in computer vision tasks.

Fully connected layer: The fully connected layer is of great importance to construct modern neural networks. It has been used in different network architectures, including in convolutional networks as the last classification layer and in multilayer perceptrons (MLPs) as the main building block. Basically, this layer implements a linear transformation that multiplies the input vector x by a matrix of learnable weights W and adds a bias vector b, giving a transformed vector:

$$y = Wx + b. \tag{1.1}$$

Nonlinear activations: While fully connected layers impose linear transformation on input



a) Nonlinear activation functions

b) Kernels learned by AlexNet

Figure 1.1 Left: nonlinear activation function used in modern neural networks. Right: visualization of the first convolutional kernels learned by AlexNet with ImageNet dataset. Both images are taken from Hadji & Wildes (2018)

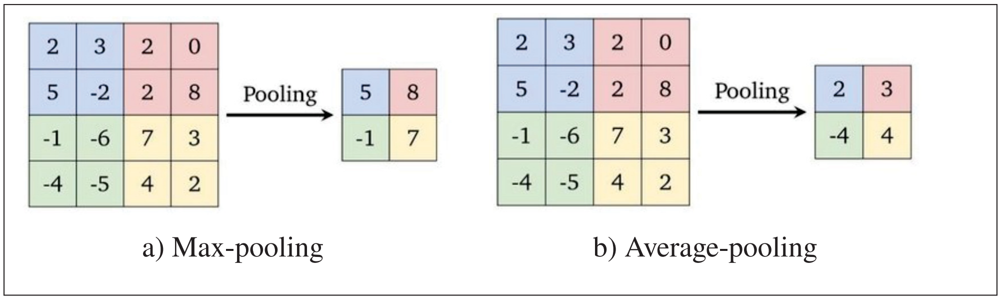
data, high-dimensional data usually exhibits nonlinearity between the inputs and outputs. An operation reflecting the nonlinear nature of such data is required. As shown in Fig. 1.1a, various nonlinear activations have been proposed for this purpose. These activations apply nonlinearity on each value independently. The logistic activation, also known as *sigmoid*, was initially used to simulate biological neurons. This function, defined as

$$\text{sigmoid}(y) = \frac{1}{1 + e^{-y}} \tag{1.2}$$

has a quasi-linear region in the middle and two saturation regions on each side. 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 rectified linear unit (ReLU) (Agarap, 2018) is a widely used activation function that prevents vanishing gradient by having a non-saturating gradient for positive values:

$$\text{ReLU}(y) = \max(y, 0) \tag{1.3}$$

Convolution: The convolution layer is one of the most important inventions in computer vision (Hadji & Wildes, 2018). This operator is inspired by neurophysiological evidence that cells in



a) Max-pooling b) Average-pooling

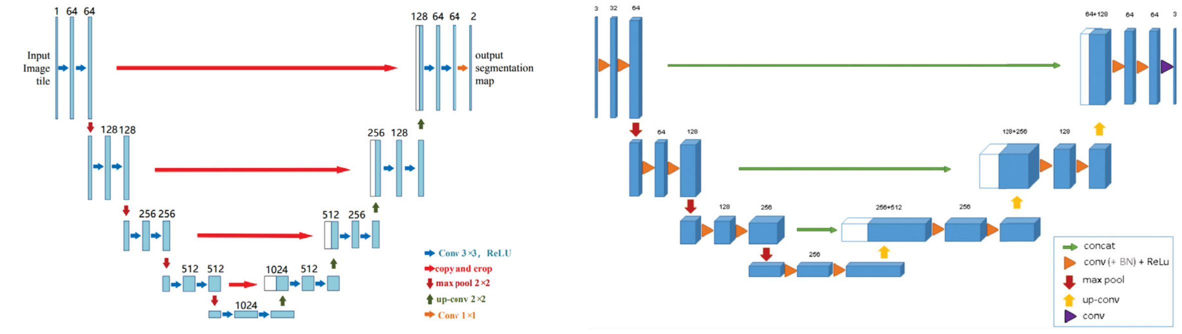
Figure 1.2 Max-pooling and average-pooling operations for a 2D image. Both image are taken from Hadji & Wildes (2018)

the retina and visual cortex are capable of detecting primitive features such as edges and bars at the early stages of visual processing. Convolution is a linear operation that *correlates* local pixel with learnable weights, also known as *kernels* or *filters*, to retrieve different patterns. A single convolution layer usually has multiple kernels, each one detecting its respective patterns from the input pixels or feature maps (Goodfellow, Bengio & Courville, 2016). Intuitively, an increased number of kernels can boost the representative ability for networks. Fig. 1.1b shows examples of kernels learned in different layers of a network trained on the ImageNet dataset. It can be observed that simple patterns such as blobs and oriented lines dominate these filters.

Pooling: Pooling is an operation in CNNs mainly used to reduce the size of feature maps and thus increase the receptive field of the network. Max-pooling, a widely used pooling method illustrated in Fig. 1.2a, returns the maximum value in the region defined by the filter size, typically set to 2×2. This can be regarded as a denoising operation where the signal with the most prominent response is preserved, while other less significant responses are discarded by reducing the spatial size. Average pooling, illustrated in Fig. 1.2b, is also widely used in deep networks, particularly for image classification. As its name suggests, the filter takes the average value of a local region instead of the maximum one.

1.2 Medical image segmentation with neural networks

Due to its broad importance in clinical applications, medical image segmentation has attracted growing attention over the years. Early methods for this task, including active-contour (Derraz, Beladgham & Khelif, 2004; Qian, Wang, Guo & Li, 2013; Chen *et al.*, 2019b; Zhang, Dong & Li, 2020a), level-set (Li, Chui, Chang & Ong, 2011; Zhang *et al.*, 2008), graph-cut (Chen, Udupa, Bagci, Zhuge & Yao, 2012; Mahapatra, 2017), normalized cut (Shi & Malik, 2000; Tang, Djelouah, Perazzi, Boykov & Schroers, 2018), and variational based methods (Han, Feng & Baciu, 2013; Freiman, Joskowicz & Sosna, 2009; Paragios, 2002) often required expert-designed features and/or slow iterative-based computation. Recently, fully-convolutional networks (FCNs) have become the most popular method for image segmentation (Saxena, 2016; Schmidhuber, 2015; Schwarz, Schulz & Behnke, 2015). These networks consist of stacks of convolutional layers, down-sampling and up-sampling blocks. Unlike neural networks designed for image classification, which gradually reduce the spatial dimension of features using pooling layers, FCNs use convolutions of different sizes followed by some nonlinear transformation (e.g., activation function, batch normalization, etc.), as well as several dimensional reduction and expansion blocks (e.g., down-sampling/up-sampling, pooling/unpooling/transposed convolution). In contrast to networks using fully-connected layers, this design enables FCNs to take an image of arbitrary size as input and predict the output labels matching the spatial resolution of the input in a differentiable way. Various networks have been proposed to improve the initial FCN network proposed by Long *et al.* (2015), such as SegNet (Badrinarayanan, Kendall & Cipolla, 2017), U-Net (Ronneberger, Fischer & Brox, 2015), PSP-Net (Zhao, Shi, Qi, Wang & Jia, 2017), and the Deep-Lab family (Chen, Papandreou, Kokkinos, Murphy & Yuille, 2017). Most of these networks adopt an encoder-decoder design. The encoder, as the first half of the network gradually reduces the spatial resolution of feature maps and acquires high-level contextual representation of the input image. On the other hand, the decoder progressively reconstructs the spatial information of the input and finally predicts as its output the probability of segmentation



a) 2D U-Net

b) 3D U-Net

Figure 1.3 Conceptual illustration of U-Net in 2D (left) and 3D (right) for medical image segmentation. Images adapted from Ronneberger *et al.* (2015) and Çiçek *et al.* (2016)

classes at each pixel.

U-Net (Ronneberger *et al.*, 2015) is one of the most popular network for medical image segmentation. This network, depicted in Fig. 1.3, adopts a symmetric architecture design comprised of five blocks for both the encoder and decoder. Each block includes several layers of convolutions with ReLU (Krizhevsky, Sutskever & Hinton, 2012) as activation function and a down-sampling or up-sampling module. This model can process the input image at different scales, thus promoting a segmentation output combining contextual and spatial information.

Another innovation of U-Net is the use of skip connections that bridge corresponding blocks of the encoder and decoder with the same spatial resolution (Ronneberger *et al.*, 2015). This particular design introduces two benefits. First, segmenting regions often requires detailed information on object boundary. A simple FCN fails to provide such high-resolution information as it is lost in the compress-reconstruction process. Skip connections solve this issue by enabling a direct flow of information from the encoder to the decoder in higher spatial resolutions. Secondly, the gradient can better pass from the decoder to the encoder with these shortcut connections, which helps update the parameters in shallow layers of the network and leads to a faster convergence. Different connectivity designs, including *Dense* (Jégou, Drozdzal, Vazquez, Romero & Bengio, 2017) and *nested* (Zhou, Siddiquee, Tajbakhsh & Liang, 2018b) skip connections were later introduced to further improve performance in some cases.

Fig. 1.4 shows the intermediate features analyzed by different blocks of a U-Net trained on a prostate segmentation dataset. Given an MRI image as the input, feature maps extracted by different blocks of the networks exhibit distinct characteristics. When going deeper in the encoder, abstract patterns with reduced resolution appear in feature maps, while feature maps from the decoder gradually delineate boundaries of the target region (the prostate in this example). In Chapter 4, we propose a novel method to explicitly regularize these intermediate features in a semi-supervised learning scenario.

U-Net was originally designed for 2D medical images but many tasks require a network to segment organs and tumours from 3D volumetric images, such as CT and MRI scans. Çiçek *et al.* (2016) proposed to extend the standard U-Net to a 3D network by replacing two-dimensional convolutions by three-dimensional ones, and reducing its depth from five to four. This drop-in replacement allows the network to aggregate information from different image axes (sagittal, coronal, and transverse) and benefits the segmentation of various structures such as the prostate (Cheng *et al.*, 2019) and brain tumours (Xu *et al.*, 2018). However, for certain datasets where the acquisition resolution is highly anisotropic, these 3D networks may actually hurt the performance (Baumgartner, Koch, Pollefeys & Konukoglu, 2017).

1.3 Loss functions for image segmentation

Neural networks acquire knowledge from data by updating their parameters to minimize a loss function (LeCun *et al.*, 2015). This function is a key factor to the performance of the segmentation network as it controls how the network corrects discrepancies between its prediction and the ground truth. Although the methods we present in the next chapters are agnostic to the choice of

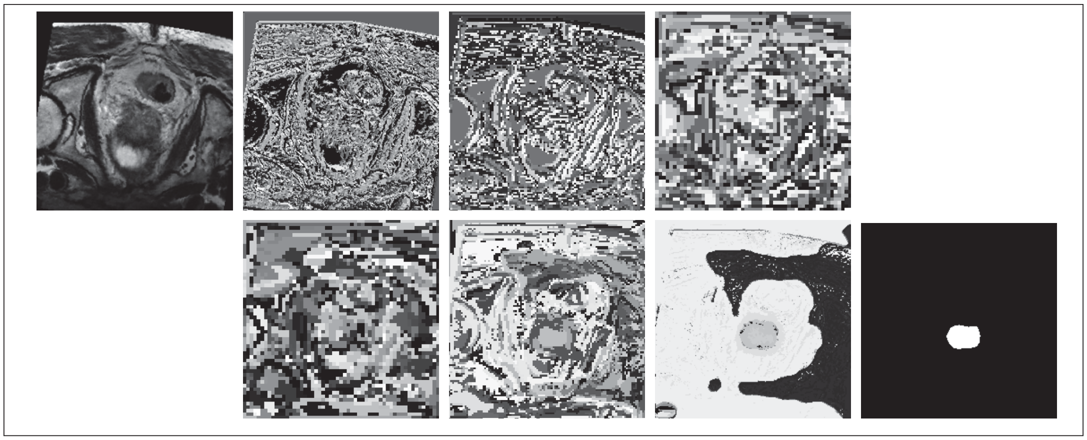


Figure 1.4 Feature activations in a well-trained U-Net. Upper: input image of a prostate slice and features generated by the first three encoder blocks. Lower: features generated by the last three decoder blocks and the output probability map for the prostate region as the foreground. Features are plotted by taking the `argmax` on the feature dimension to reflect the activated feature class number. Each intensity value thus corresponds to different feature dimension.

supervised loss function (i.e., the loss computed on labeled examples), we present some of them as they are an important component for medical image segmentation.

Given a training dataset $\mathcal{D}_{\text{train}} = \{(x_i, y_i)\}_{i=1}^{N}$ with N image and ground truth pairs, where $x_i \in \mathbb{R}^{H\times W}$ denotes the i^{th} image (unlike natural images, medical images are typically encoded in gray-scale: $1 \times H \times W$ and we ignore the first dimension for the sake of simplification), and $y_i \in \{1, \ldots, K\}^{H\times W}$ denotes the densely annotated ground truth with K classes. During optimization, we want the segmentation network to learn an approximation of the nonlinear function $f_\theta(\cdot)$, parameterized by θ, which maps an image x_i to its labels y_i. Considering the network's output $s_i = f_\theta(x_i) \in \Delta^{H\times W\times K}$ as a dense prediction distribution indicating the probability of assigning the class k $(1 \leq k \leq K)$ to each pixel, training a neural network consists in finding the parameters θ^* minimizing a loss function on $\mathcal{D}_{\text{train}}$:

$$\theta^* = \arg\min_{\theta} \; \mathcal{L}(\theta) = \frac{1}{N}\sum_{i=1}^{N} l(s_i, y_i) \tag{1.4}$$

where $l(s_i, y_i)$ is a penalty comparing the prediction s_i for image x_i and the corresponding ground truth y_i. $\Delta^{H\times W\times K}$ refers to simplex of probabilities which sums to one along axis K. A standard requirement for penalty function l is that it must be continuous and differentiable. When there is no ambiguity, in what follows, we simplify the notation by dropping the index i of a training example.

Several losses were proposed for image segmentation. Cross-entropy (CE) (Yi-de, Qing & Zhi-Bai, 2004) and its variants are among the most widely used ones. Adapted from classification tasks, the CE loss for segmentation considers each pixel in an input image independently, and computes penalty for an image as

$$l(s, y) = -\frac{1}{N \times W} \sum_{p=1}^{H\times W} \sum_{k=1}^{K} \mathbf{1}(y^p = k) \log(s_k^p). \tag{1.5}$$

In this equation, y^p denotes the true class label at pixel position p $(1 \leq p \leq H\times W)$ and s_k^p is the probability of assigning class k $(1 \leq k \leq K)$ at the same position. $\mathbf{1}(\cdot)$ is an indicator function which returns 1 if the input expression is true. However, since medical image segmentation can suffer from severe class imbalance (e.g., the background region may dominate foreground ones), using cross entropy may result in a significant bias where the network mostly focuses on the background class and ignores foreground ones. Weighted cross-entropy (Pihur, Datta & Datta, 2007) mitigates this problem by instead weighting each class separately based on the marginal distribution of each class. The idea is simple: if a class appears more often than other ones, we can multiply the loss for this class by a small weighting coefficient to reduce its contribution. As expressed in Equ. 1.6, the weighted CE loss assigns a weighting coefficient w_k to each class, which is often inversely proportional to its class frequency: $\frac{H\times W}{\sum_p^{H\times W} \mathbf{1}(y^p=k)}$.

$$l(s, y) = -\frac{1}{N \times W} \sum_{p=1}^{H\times W} \sum_{k=1}^{K} w_k \cdot \mathbf{1}(y^p = k) \log(s_k^p) \tag{1.6}$$

As an alternative to the weighted CE loss, *Dice* loss (Sudre, Li, Vercauteren, Ourselin & Cardoso, 2017) addresses the class imbalanced problem by mimicking the famous DSC coefficient (Zou *et al.*, 2004). This loss maximizes the overlap between the ground truth and predicted regions, as follows:

$$l(s, y) = 1 - \frac{2 \times \sum_{p=1}^{H\times W} \sum_{k=1}^{K} \mathbf{1}(y^p = k) \cdot s_k^p + \epsilon}{\sum_{p=1}^{H\times W} \sum_{k=1}^{K} \mathbf{1}(y^p = k) + s_k^p + \epsilon} \tag{1.7}$$

where ϵ is a small constant to avoid numerical instability. Recent work (Liu, Dolz, Galdran, Kobbi & Ayed, 2021a) showed the Dice loss to have intrinsic bias towards small structures, without the need to a specific weighting mechanism, making it one of the most popular losses for medical image segmentation. However, it can result in a slower convergence compare to the CE loss. The combo loss (Taghanaki *et al.*, 2019), which combines Dice and CE, offers the advantages of both these losses.

Since the Dice and CE losses do not give more importance to predictions at the boundary of regions, they may produce segmentations with a high overlap accuracy but a correspondingly low contour accuracy. The boundary loss (Kervadec *et al.*, 2019a), which uses a signed distance map to estimate the integral of distances between a predicted contour and the ground truth one, has been proven effective for segmentation tasks imbalanced data. The boundary-aware segmentation method in (Karimi & Salcudean, 2019) estimates the Hausdorff distance (HD) between the predicted and ground truth contours using a distance transform and a morphological erosion operation. More recently, Yang *et al.* (2020) proposed a loss based on Laplacian filter that increases the importance of pixels near the boundary.

1.4 Learning with reduced supervision

Methods reducing the need for pixel-wise annotations have been intensively investigated in the literature. These methods can be roughly divided in the following three categories:

- Methods using *weak annotations*. Weak annotations vary from an application to another, and include *image-level tags* that denote the presence or absence of a specific class in the image (e.g., tumor), *points or scribbles* that indicate the partial (incomplete) location of a region of interest, or *bounding boxes* within which a region of interest is contained. These weak annotations can sometimes be obtained automatically from clinical reports, or require less manual works from experts compared to the dense annotation of images. However, they only provide partial spatial information on segmentation targets.

- Methods employing *anatomical priors*. To leverage incomplete supervision signals effectively, weak annotations are often combined with *domain knowledge priors* (El Jurdi, Petitjean, Honeine, Cheplygina & Abdallah, 2021). In clinical settings, medical images are usually acquired with strict protocols, and anatomical regions to segment often exhibit strong similarity across patients. These anatomical priors, which can be obtained beforehand, can encode a broad range of information, such as constraints on the size or position of a target region or some global property such as connectivity and boundary smoothness. These priors are used to guide the training of a network so that it outputs more plausible segmentations on images without dense annotations. Such priors can also be used in full supervision to further regularize the segmentation predictions (Mirikharaji & Hamarneh, 2018; Karimi & Salcudean, 2019).

- Methods exploiting *unlabeled images*. Unlabeled medical images are more abundant and easier to obtain than those with annotations. These images can be leveraged to boost both the accuracy and the generalization capability of a network, especially when combined with a few densely-annotated images. The main challenge of this approach is designing an effective way to exploit unlabeled images. A strategy used by many methods, including data augmentation, co-training, generative models, adversarial training, and consistency-based methods, is to design a regularization loss on unlabeled samples and optimize it jointly with a supervised loss defined on labeled data. Recently, representation learning has emerged as a powerful way to exploit the readily-available knowledge contained in unlabeled data. In this approach, a network is first pre-trained on *all* unlabeled data with a given proxy task, and then fine-tuned on the target task using only a small set of labeled data.

Methods in these three categories are complementary, and they can be used in a collaborative way to improve performance when labeled data in scarce. In the following subsections, we present recent works applying such methods in the context of medical image segmentation.

1.4.1 Methods using weak annotations

In weakly supervised learning, partial or incomplete annotations are used to train the segmentation network. Multiple instance learning (MIL) is a well-known approach for exploiting *image-level* annotations. In this approach, an image is considered as a bag of instances (individual pixels or patches), and a bag is considered as positive if at least one instance includes the object of interest, otherwise it is negative. Traditional MIL algorithms usually consist of three steps: network initialization, bag-based update, and object-relocalization. Since the last two steps are performed iteratively, MIL can be sensitive to network initialization and can get stuck in sub-optimal solutions during the iterative optimization. Pathak, Shelhamer, Long & Darrell (2014) proposed one of the first MIL frameworks for natural image segmentation using FCNs. This frameworks considers the highest activated location in a score heatmap, with image tags as the supervision signal. In the medical domain, MIL was used on microscopy and histopathological images. Kraus, Ba & Frey (2016) proposed a new pooling layer to aggregate the bag's prediction for breast cancer screening. In Lerousseau *et al.* (2020), highly-confident patches are selected to update the neural network for tumor segmentation. Recently works using *image-tags* also leverage class activation maps (CAMs) (Zhou, Khosla, Lapedriza, Oliva & Torralba, 2016) to obtain localization cues for segmentation. The core idea of this approach is that the high-level feature maps of a fully-trained classification network contain both semantic and spatial information for each class. Since CAMs represent the most activated regions for a given image tag, it can be used to initialize the training of a segmentation network. The resulting prediction can then refined with a smoothness prior, for example using a conditional random field (CRF) (Chen *et al.*, 2017). Wei *et al.* (2017); Sun, Shi, Zhang & Huang (2021) further improved the localization ability of CAMs by adopting adversarial erasing for natural images, the most discriminative part of images/features are removed. In medical imaging, Schumacher, Genz & Heinrich (2020) proposed using CAMs to segment the pancreas in CT scans, while Patel & Dolz (2021) enhanced CAMs with transformation equivariance constraints to segment brain tumors and prostates in multi-modal MRI.

Weak labels in the form of points or scribbles are also popular for medical segmentation tasks as they considerably reduce the annotation effort. Qu *et al.* (2019) used point supervision together with a clustering assumption to segment histopathology images. Specifically, they employed a pixel-wise cross-entropy loss to train the network, and a clustering strategy to generate the pseudo labels. Roth, Yang, Xu, Wang & Xu (2021) proposed using multiple points in an interactive way as supervision. Combined with random walk regularization, this approach led to accurate multi-organ segmentation in abdominal CT images. Kervadec *et al.* (2018) employed partial cross entropy with simple region size constraints to train a segmentation network on scribbles with a few annotated foreground pixels.

Bounding boxes are simple annotations that indicate the location of a target region using the tightest rectangle that covers this region. It reduces the annotation burden by requiring only two extreme points, e.g., the upper-left and lower-right corners of the box. A common way of performing segmentation given a bounding box consists of two iterative steps: pseudo-label proposal generation and network update. DeepCut (Rajchl *et al.*, 2017) proposed to generate reliable pseudo-labels using a CRF prior, while BoxSup (Dai, He & Sun, 2015) generated pseudo-labels via an unsupervised region proposal algorithm. Recently, Hsu, Hsu, Tsai, Lin & Chuang (2019) presented an MIL approach with tightness priors to optimize a segmentation network. In their method, pixels in different crossing lines within the box are used as positive bags, while those in lines outside the box are considered as negative bags. Wang & Xia (2021) extended this approach by also considering rotated crossing lines within the box as positive bags. To improve the weakly-supervised segmentation of medical images, Kervadec *et al.* (2020) proposed an alternative method that constrains the softmax prediction with similar tightness and size priors.

1.4.2 Methods using anatomic priors

Medical segmentation can also leverage anatomical knowledge to effectively regularize the network prediction. These constraints, which are often available beforehand, can include the approximate size range of a target region, the shape prior of this region, or properties like convexity or connectivity. Pathak, Krahenbuhl & Darrell (2015) first proposed using inequality constraints on region size to regularize the output distribution of a network for segmenting natural images. The proposed approach added a latent distribution satisfying the constraints and used a Lagrangian dual method to enforce similarity between network prediction and this latent distribution. The optimization in this method is based on a two-step alternating update scheme: the neural network is updated by gradient descent while the latent space by dual method. Kervadec *et al.* (2018); Kervadec *et al.* (2019b) proposed a similar idea to constrain the output size under inequality constraints for organ segmentation with MRI images. In their method, the authors approximated the size of a organ as the soft sum of all probabilities for foreground classes and imposed an L_2 loss function to penalize sizes that do not satisfy a pre-defined range. However, using the sum of foreground probabilities to estimate the size can be inaccurate when probabilities are not binary. Further work (Bateson, Kervadec, Dolz, Lombaert & Ayed, 2019) extended the size prior in a domain adaptation scenario, where the image appearance changes across modalities but the anatomical size remains the same. In a different approach, Zhang, Zhong & Li (2020b) used the adversarial training of a discriminator to learn segmentation constraints directly from the data.

Integrating shape knowledge in segmentation networks has also attracted a lot of attention in medical imaging. Zotti, Luo, Lalande & Jodoin (2018) first proposed using a shape atlas in a FCN for cardiac segmentation in Cine MRI images. In their method, the shape prior is defined as a *probability map* with values ranging from 0 to 1, measuring the pixel-wise empirical proportion of each class based on ground truth labels. An $\mathcal{L}_2$ loss was used to steer the output prediction towards this statistical shape prior. A shape-based method was also proposed by Huang *et al.* (2021) for liver and spleen segmentation in abdomen CT images. In this method, a deep atlas prior (DAP) is used to combine the probabilistic prior with an uncertainty prior. Moreover, Simantiris & Tziritas (2020) proposed to explicitly penalize pixels identified as anatomically impossible. The star-shape prior is another interesting prior for certain tasks such as skin lesion segmentation. Mirikharaji & Hamarneh (2018) proposed a differentiable loss modeling this prior. The proposed loss imposes that, for pixels predicted as foreground, all pixels on a line segment from this pixel to the foreground region center are also predicted as foreground. Other works exploited complementary information derived from pixel-wise labels to guide the segmentation prediction. In addition to predicting the segmentation probabilities, the networks proposed by Li, Zhang & He (2020b) and Xue *et al.* (2020) also predicts a signed distance map (SDM) measuring the distance of pixels to their nearest point on the foreground region boundary (a negative value means that the pixel is inside the foreground region). Navarro *et al.* (2019) proposed a multi-task learning approach to enhance segmentation by jointly predicting a segmentation mask, a SDM and a contour map. In contrast, Cheng *et al.* (2020) predicts a direction field that evaluates for each pixel the direction opposite to the nearest point on the foreground region boundary. This field is used to “rectify” the features of pixels near the boundary, which have a higher uncertainty.

Other priors such as connectivity and convexity were also explored to guide the segmentation prediction (Shi & Li, 2021; Ganaye, Sdika, Triggs & Benoit-Cattin, 2019). However, these prior require to formulate an explicit regularization loss, which may be difficult in some cases. Recently, it was shown that high-order anatomical prior can be acquired *automatically* from an auto-encoder (AE) network that learns a compact manifold of feasible predictions from a set of densely-annotated images. Oktay *et al.* (2017) showed that a well-trained AE can help the network produce more anatomically-plausible segmentations. Painchaud *et al.* (2020) further leveraged a constrained variational AE to refine the prediction of a segmentation model. Compared to other methods, their approach guarantees the validity (i.e., satisfaction of anatomical constraints) of refined segmentations.

1.4.3 Semi-supervised learning methods

Various semi-supervised methods have been proposed to boost the accuracy and generalization of deep learning models using a small set of unlabeled images and a larger set of unlabeled ones. The majority of these methods can be grouped in five main categories based on: consistency, co-training based, information-metrics, adversarial learning or representation learning.

Consistency-based methods

Consistency-based methods are among the most popular semi-supervised techniques for segmentation. The core idea of these methods is to regularize the learning of a network using unlabeled images, based on the well-known transformation-invariant principle that the network's prediction for a given image should be invariant to information-preserving transformations applied to this image. The Π-model (Laine & Aila, 2016) was among the first to tackle semi-supervised learning in classification. This model enforces consistent predictions for two perturbed versions of the same unlabeled image. Based on a similar idea, Bortsova *et al.* (2019) proposed to enforce transformation equivalence (or *equivariance*) on a segmentation network $f(\cdot)$, i.e. $f(T(x)) = T(f(x))$ where T is a random geometric transformation (e.g., image rotation, scaling, cropping, etc.). Using this approach, the authors successfully boosted the accuracy of a network for the semi-supervised segmentation of chest X-ray images. Enforcing consistency with adversarial examples is another promising way to improve the robustness of models in semi-supervised learning settings (Miyato, ichi Maeda, Koyama, Nakae & Ishii, 2015). Instead of using a set of predefined transformations, this approach transforms a training image to maximize the divergence of the network's prediction. The well-known Mean Teacher (Tarvainen & Valpola, 2017) method maintains a separate Teacher network whose parameters are the exponential moving average of the Student's. An unsupervised loss is imposed to minimize the prediction discrepancy between the Teacher and the Student on unlabeled examples. This method, which distils knowledge over different training iterations to generate reliable pseudo-labels for unlabeled images, is widely used in both semi-supervised classification (Tarvainen & Valpola, 2017) and segmentation (Perone & Cohen-Adad, 2018). It is at the core of semi-supervised segmentation algorithms for various medical tasks, including skin lesion, liver, retinal fundus (Li *et al.*, 2020c), and left atrium (Yu, Wang, Li, Fu & Heng, 2019) segmentation.

The concepts of uncertainty (Yu *et al.*, 2019), multitask learning (Luo, Chen, Song & Wang, 2020), as well as global or local information distillation (Hang *et al.*, 2020) were also proposed to further boost the performance of consistency-based methods for semi-supervised segmentation. In a related problem called unsupervised domain adaptation (UDA), their is also a shift in the distribution of unlabeled images compared to labeled ones (e.g., labeled images can be MRI whereas unlabeled ones are CT). This domain shift can prevent a segmentation network to generalize well on out-of-domain data. To mitigate this issue, Li, Wang, Yu & Heng (2020a) proposed a dual teacher approach in which the first teacher aims to learn cross-domain knowledge with a Cycle-GAN, while the other seeks to increase in-domain generalization.

Co-training based methods

Instead of using the consistency of a single network for different image transformations, co-training exploits the prediction consistency of several models with diversified views as their respective input. The first co-training algorithm (Blum & Mitchell, 1998) was proposed for web page classification by employing two independent views: the description of a web page and the words in hyperlinks that point to that page. In this algorithm, two models are trained with different views and the most confident unlabeled sample from one view is considered as labeled and contributes to the training of the other view. Medical images such as CT and MRI scans can often be split into three views corresponding to *sagittal*, *coronal*, and *transverse* (axis) planes. Xia *et al.* (2020) proposed to train three segmentation networks extended from a 2D model, each of them having a different view of the data as input. Reliable pseudo-labels can be created by merging the three independent predictions for the same unlabeled image, while also considering prediction uncertainty. Likewise, Huang, Zheng, Hu, Zhang & Li (2020) used co-training to train a unified framework that learns from multiple few-organ datasets.

Information-metric based methods

Information-based metrics are widely employed in semi-supervised learning to regularize predictions for unlabeled examples. One of the most popular metrics for this task is *entropy*, which measures the uncertainty in a system. A high value of entropy means that the uncertainty (randomness) in the system is high and it is difficult to predict its states. On the other hand, if entropy is low, predicting its state is easier. The intuition of reducing entropy in semi-supervised learning is that making predictions more confident for unlabeled samples implicitly pushes the decision boundary away from these samples. Following the clustering assumption, decision boundaries are more likely to lie in low-density regions than in high-density ones. Grandvalet & Bengio (2005) demonstrated the effectiveness of entropy minimization in image classification. Vu, Jain, Bucher, Cord & Pérez (2019) then applied this principle to image segmentation. Entropy minimization has also been used to improve the generalization of neural networks for medical segmentation tasks such as left atrium (Hang *et al.*, 2020), leukocyte (Wu, Fan, Zhang, Lin & Li, 2021a) and lower spine (Bateson, Kervadec, Dolz, Lombaert & Ayed, 2020) segmentation.

Mutual information (MI) is an abstract information metric which measures the mutual dependence between two random variables. The concept of MI is intimately linked to entropy, as MI between two random variables $\mathcal{X}$ and $\mathcal{Y}$ can be expressed as difference between the entropy of $\mathcal{Y}$ and its entropy conditioned on $\mathcal{X}$:

$$I(\mathcal{X};\mathcal{Y}) = H(\mathcal{Y}) - H(\mathcal{Y}|\mathcal{X}). \tag{1.8}$$

Typically, we seek to maximize the MI between X an Y, where $\mathcal{X}$ is the input of a neural network and $\mathcal{Y}$ corresponds to its predicted probability distribution. The first term (marginal entropy, $H(\mathcal{Y})$) then encourages $\mathcal{Y}$ to be uniformly distributed, which avoids trivial solutions where all predictions in $\mathcal{Y}$ are assigned to only a few classes. On the other hand, the second term (conditional entropy, $H(\mathcal{Y}|\mathcal{X})$) enforces the network to output confident predictions. Because of these properties, MI maximization is often used for image clustering and unsupervised segmentation tasks (Ji, Henriques & Vedaldi, 2019). Based on this idea, Zhao, Wang, Yang & Cai (2019b) proposed a regional MI loss to consider high-order dependencies between adjacent pixels. However, this loss was tailored for supervised learning and did not exploit unlabeled images. In medical imaging, Xi (2019) used MI in collaboration with generative models for brain tumor segmentation.

Adversarial learning based methods

Adversarial based methods became popular in semi-supervised segmentation thanks to the invention of generative adversarial networks (GANs) in 2014 by Goodfellow *et al.* (2014). The learning framework of GANs typically consists of two components:

1. A generator $G(z)$, which is a differentiable function that maps z from a prior distribution $P(z)$ to data space. The generator's goal is to produce samples that match the distribution of real data.
2. A discriminator $D(x)$, which tries to tell whether the input image x comes from real data or was generated by $G(\cdot)$.

The generator and discriminator are trained alternatively in a min-max game style. The generator $G(\cdot)$ is trained to output realistic samples so that it can fool the discriminator $D(\cdot)$, while the discriminator is optimized to separate the real samples from the generated ones. It was shown in (Goodfellow *et al.*, 2014) that this competitive training minimizes the *Jensen-Shannon divergence* (JSD) between the true data distribution and the distribution produced by the generator $G(\cdot)$. Springenberg (2015) was the first to use GANs for semi-supervised learning. In their algorithm, the discriminator is no longer a binary classifier but a K-way classifier. When given an unlabeled image, the discriminator tries to give a confident prediction to one of the K classes (lower conditional entropy). Conversely, when a fake image arrives, the discriminator gives a uniform prediction (high conditional entropy). Odena (2016) further improved the method by increasing K to $K + 1$ classes with *Fake* as the added class. In medical image segmentation, GANs are widely used as a data augmentation technique for synthesizing new labeled images. Zhao, Balakrishnan, Durand, Guttag & Dalca (2019a) proposed to learn a generative network to synthesize new brain MRI scans from both labeled and unlabeled samples using two complimentary networks, a spatial deformation network and appearance transform network. Cycle-GANs (Zhu, Park, Isola & Efros, 2017) are also commonly used in semi-supervised learning. This approach closes the appearance gap between images from different domains so that a segmentation network can be trained with more (synthesized) labeled data. Another branch of research considers the segmentation network itself as the generator and only introduces a discriminator to guide the segmentation network toward plausible predictions. In Zhang *et al.* (2017c), a discriminator network should distinguish between the segmentation of labeled and unlabeled images. To fool the discriminator, the segmentation model must therefore perform as well on unlabeled images as on labeled ones. Xue, Xu & Huang (2018) further boosted the performance of this approach for skin lesion segmentation with a multi-scale discriminator.

Representation learning based methods

Methods based on representation learning adopt a different strategy to exploit unlabeled data. While other approaches typically design a loss on unlabeled data, optimized jointly with a supervised loss, methods falling in this category instead *pre-train* the network on a pre-defined task using only unlabeled data. In a second (separate) stage, the pre-trained network is then fine-tuned using a few labeled sample. The main hypothesis of this strategy is that the unlabeled pre-training task helps the network learn a useful feature representation that can improve its accuracy and generalization ability on the target task.

Representation learning, which can also be referred to as self-supervised learning, differs from transfer learning where a network is pre-trained from a large-scale annotated dataset such as ImageNet (Deng *et al.*, 2009). Hosseinzadeh Taher, Haghighi, Feng, Gotway & Liang (2021) showed that self-supervised methods generally outperform those based on transfer learning based for both classification and segmentation tasks with medium sized medical datasets. Various pre-defined tasks, called pretext tasks, have been proposed for medical image segmentation. Chen *et al.* (2019a); Bai *et al.* (2019) pre-trained the network to predict the relative position of different image patches. Chen *et al.* (2019a) instead proposed to restore the context details given deteriorated images augmented by cutout (Devries & Taylor, 2017). Taleb, Lippert, Klein & Nabi (2021) applied jigsaw puzzle solving on multiple MRI modalities. The Models Genesis (Zhou, Sodha, Pang, Gotway & Liang, 2021) method restores an original volumetric image from a degraded version transformed using Bezier curve warping, pixel shuffling and cutout.

Contrastive learning (Oord, Li & Vinyals, 2018) has emerged as a powerful pre-training approach to boost classification or segmentation when using unlabeled data. In this approach, a network is pre-trained to perform *instance-based* classification. The core idea of this approach is simple: pulling close to each other the representations of an image under different transformations (or degradations), and pushing away those from different images despite of their true label classes. Performing contrasting learning in such a way leads to significant improvement when using the learned representations with a simple linear classifier (probe) or a K-nearest-neighbor protocol. As shown by Chen, Fan, Girshick & He (2020c), in some cases, the performance of this simple classification approach on downstream tasks can surpass the one obtained with conventional full supervision. In this work, the authors studied the benefits of contrastive learning in a semi-supervised classification scenario where a large network is pre-trained with contrastive learning and then fine-tuned using a small fraction of labeled data. To further improve the downstream task, knowledge distillation was also adopted to train a small network under the guidance of the fine-tuned one. The authors showed the classification accuracy of their method, using only 10% of labeled data, to be on par with full supervision.

So far, the application of contrastive learning to medical image segmentation has been limited. Chaitanya *et al.* (2020) proposed to pre-train the encoder and decoder of a segmentation network. The encoder is trained with a contrastive loss that maximizes the cosine similarity of representations for 2D images (slices) from the same subject or in the same position of a 3D volume. On the other hand, the decoder is trained to output dense features that are invariant to geometric and intensity transformations. Zeng *et al.* (2021) further improved the encoder pre-training by considering the distance between pairs of 2D slices in the contrastive loss. Despite the improvement brought by these methods in scenarios where labeled data is very limited, how to effectively leverage rich meta-information carried by medical images for proxy tasks and how to pre-train dense feature maps are still open questions.

1.5 Discussion

In this chapter, we presented the necessary background knowledge to understand the context of this thesis, including the main neural network architectures employed in medical image segmentation and their optimization losses. We also highlighted the challenges of segmenting medical images, in particular the scarcity of densely-annotated data, and described the main approaches proposed to train deep segmentation networks with reduced supervision.

Despite the significant progress made in recent years to improve medical image segmentation in scenarios where labeled data is scarce, current approaches still suffer from important limitations that impede their use in real-life clinical applications. For instance, while anatomical constraints have been used as prior to improve segmentation in a weakly-supervised setting, these constraints are typically formulated in a continuous optimization problem (Kervadec *et al.*, 2019b,a). However, many constraints on a target region, such as size or boundary length, are discrete in nature and tackling them in a discrete optimization setting can lead to a better solution. Moreover, although co-training has been used successfully for semi-supervised medical segmentation (Xia *et al.*, 2020; Huang *et al.*, 2020), previous methods based on this idea require to have separate views (e.g., different imaging planes of a volumetric image) as input to the different models, which may not be possible in all applications (e.g., when the resolution along a dimension of the volumetric image is low). Furthermore, existing segmentation approaches based on mutual information maximization also have three important limitations. First, they only consider the output of the network Ji *et al.* (2019); Zhao *et al.* (2019b) and ignore the relationship of dense features inside intermediate layers of a deep network. Second, they are typically used for unsupervised segmentation, and their usefulness for semi-supervised segmentation tasks has not been fully explored. Third, since these methods do not consider boundary cues such as edges in an image, the spatial clusters learned by maximizing MI may not correspond to actual anatomical structures in that image. Finally, while contrastive learning approaches for segmentation have led to substantial improvements when very few labeled images are available, these approaches rely on noisy meta-labels (e.g., the arbitrary grouping of 2D slices in a volumetric image), as shown by Chaitanya *et al.* (2020), which can lead to a poor representation. The following chapters of the thesis present novel semi-supervised learning methods for medical image segmentation that address these limitations.

CHAPTER 2

DISCRETELY-CONSTRAINED DEEP NETWORK FOR WEAKLY SUPERVISED SEGMENTATION

Jizong Peng[1] , Hoel Kervadec[2] , Jose Dolz[1] , Ismail Ben Ayed[2] , Marco Pedersoli[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] Department of Automated Production, École de Technologie Supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3

Paper published in *«Neural Networks»* in July 2020.

2.1 Presentation

This chapter presents the article “Discretely-constrained Deep Network for Weakly Supervised Segmentation" by Peng, Kervadec, Dolz, Ben Ayed, Pedersoli and Desrosiers accepted to the journal *Neural Networks* for publication on 18 July 2020. The initial results of this paper were presented as a poster paper in *Medical Imaging Meets NeurIPS*. The objective of this research is to propose a novel method, based on the alternating direction method of multipliers (ADMM) algorithm, to train a CNN with discrete constraints and regularization priors. With our proposed method experimentally tested in weakly-supervised segmentation settings, we have achieved significant improvements compared to existing approaches in terms of segmentation accuracy, constraint satisfaction and convergence speed.

2.2 Introduction

Semantic image segmentation is a fundamental problem in computer vision, which requires assigning the proper category label to each pixel of a given image. This problem is essential to various applications such as autonomous driving (Luc, Neverova, Couprie, Verbeek & LeCun,

2017) and neuroimaging (Yuan, Chao & Lo, 2017; Dolz, Desrosiers & Ben Ayed, 2018a). Among the wide range of segmentation systems, approaches based on deep convolutional neural network (CNN) have recently attracted great attention (Litjens *et al.*, 2017). While providing state-of-art performance in many segmentation tasks, CNN-based approaches usually require a large training set of fully-annotated images that can be both time-consuming and expensive to obtain. This limitation is especially important in medical imaging where access to patient data is restricted and annotation requires expert-level knowledge. This impedes the development of clinical tools which necessitate an accurate, semi-automated delineation of target regions, including various anatomical structures and types of lesion like tumors.

Unlike fully-supervised approaches for segmentation, which require every pixel of training images to be labeled, weakly-supervised methods can learn with partially-labeled images or with uncertain/noisy labels. Since weak annotations such as image-level tags (Kervadec *et al.*, 2019b; Pathak *et al.*, 2015), bounding boxes (Dai *et al.*, 2015; Rajchl *et al.*, 2017) and scribbles (Kervadec *et al.*, 2019b; Lin, Dai, Jia, He & Sun, 2016) are easier to obtain than expensive pixel-wise annotations, such methods have been intensively researched for segmentation. Another popular strategy for weakly-supervised segmentation uses prior knowledge in the form of constraints to guide the segmentation of unlabeled or partially-labeled images during training (Kervadec *et al.*, 2019b; Pathak *et al.*, 2015). This is particularly useful for medical segmentation problems, where information about the target region is often known beforehand. So far, methods proposed for constrained segmentation have exploited continuous optimization techniques where binary variables of the discrete solution space are relaxed to the continuous [0, 1] interval representing the class probabilities of pixels. Yet, many constraints such as bounds on the size of segmented regions are better expressed discretely. For example, in continuous methods, the size of a region is typically estimated by computing the sum of pixel probabilities corresponding to this region. Limiting this sum does not however guarantee that the actual size of the region will meet constraints once the probabilities are thresholded. Similarly, regularization priors for segmentation like boundary length (Boykov, Veksler & Zabih, 2001), star-shapeness (Veksler, 2008) and compactness (Dolz, Ayed & Desrosiers, 2017a) are usually discrete, and optimizing them in a continuous framework is susceptible to local minima. Lastly, solving sub-problems in a discrete manner, instead of using gradient descent, benefits from globally-optimal algorithms which can significantly improve the current solution in a single update step.

In this paper, we address the limitations of existing approaches for weakly-supervised constrained segmentation by exploring this problem from a discrete perspective. The main contributions of our work are the following:

- We present a first method to train a CNN with discrete constraints and regularization priors. The proposed method uses an efficient strategy, based on the alternating direction method of multipliers (ADMM) algorithm, to separate the optimization of network parameters with SGD from optimizing discretely-constrained segmentation labels. We show that updating these discrete labels can be done in polynomial time with guarantee of solution optimality.
- We apply the proposed method for the segmentation of medical images with weak annotations. While previous works have considered either size constraints (Kervadec *et al.*, 2019b; Pathak *et al.*, 2015) or boundary length regularization (Bai *et al.*, 2017) for segmentation, our method combines these two priors in a single efficient model. Experiments on three segmentation problems show our method to yield significant improvements compared to existing approaches in terms of segmentation accuracy, constraint satisfaction and convergence speed.

The rest of this paper is organized as follows. In Section 2.3 we give an overview of related work. We then present our discretely-constrained model in Section 2.4 and describe the experiments to evaluate its performance in Section 2.5. Finally, we give experimental results in Section 4.6 and draw conclusions in Section 2.7.

2.3 Related work

2.3.1 Constrained segmentation

Constrained optimization has played a key role in various application domains, such as control theory (Sun, Mou, Qiu, Wang & Gao, 2018; Qiu, Sun, Wang & Gao, 2019). Several works have tackled the problem of weakly-supervised segmentation by imposing constraints on deep CNNs (Kervadec *et al.*, 2019b; Pathak *et al.*, 2015; Jia, Huang, Eric, Chang & Xu, 2017; Zhou *et al.*, 2019a). In Pathak *et al.* (2015), Pathak et al. propose a latent distribution and KL-divergence to constrain the output of a segmentation network. Their method allows decoupling the optimization of network parameters with stochastic gradient descent (SGD) from the update of the latent distribution under constraints. It is used in a semi-supervised setting to impose size constraints and image-level tags (i.e., force the presence or absence of given labels) on the regions of unlabeled images. Moreover, a simple L_2 penalty term was proposed in (Jia *et al.*, 2017) to impose equality constraints on the size of the target regions in the context of histopathology image segmentation. More recently, Kervadec et al. (Kervadec *et al.*, 2019b) showed that imposing inequality constraints on size directly in gradient-based optimization, also via an L_2 penalty term, provided better accuracy and stability than the approach in (Pathak *et al.*, 2015) when few pixels of an image are labeled. Similarly, Zhou et al. (Zhou *et al.*, 2019a) embedded prior knowledge on the target size in the loss function by matching the probabilities of the empirical and predicted output distributions via the KL divergence. As directly minimizing this term by standard SGD is difficult, they proposed to optimize it by using stochastic primal-dual gradient. While these works have helped improve segmentation in a weakly-supervised setting, they have mainly focused on continuous optimization methods. In our work, we show that this problem can be solved more efficiently with discrete optimization.

2.3.2 Discrete-continuous optimization

Discrete optimization has a long history in research, playing a key role in various problems of computer science and applied mathematics (Wolsey & Nemhauser, 2014). Theoretical results in this field, like the minimization of submodular functions (Cunningham, 1985), have given us efficient tools to solve complex decision problem involving discrete variables. In recent years, significant efforts have been made to combine these powerful tools with optimization techniques for continuous problems. In Miksik, Vineet, Pérez, Torr & Sévigné (2014), an ADMM algorithm is used to perform distributed inference in large-scale Markov Random Fields (MRF), using both discrete and continuous variables in the optimization. A similar idea is proposed in Dolz, Ben Ayed & Desrosiers (2017b) to minimize discrete energy functions (submodular and non-submodular) for distributed image regularization. Other lines of work include discrete-continuous methods based on ADMM for incorporating high-order segmentation priors on the target region's histogram of intensities (Karnyaczki & Desrosiers, 2015) or compactness (Dolz *et al.*, 2017a). More recently, similar techniques have been proposed to include regularization priors directly in the learning process (Laude *et al.*, 2018; Marin, Tang, Ayed & Boykov, 2019). Despite showing promising results, none of these works have focused on constrained segmentation. To our knowledge, our work is the first employing a discrete-continuous framework for weakly-supervised segmentation with constraints.

2.4 Methodology

In this section, we first provide a formal definition of the weakly-supervised segmentation problem considered in our work. We then give an overview of penalty-based and ADMM-based methods for solving constrained optimization problems, highlighting the advantages of the latter. Last, we present our ADMM-based method for segmentation with regularization prior and size constraints, and analyze its computational complexity. To facilitate the reading of this section, a

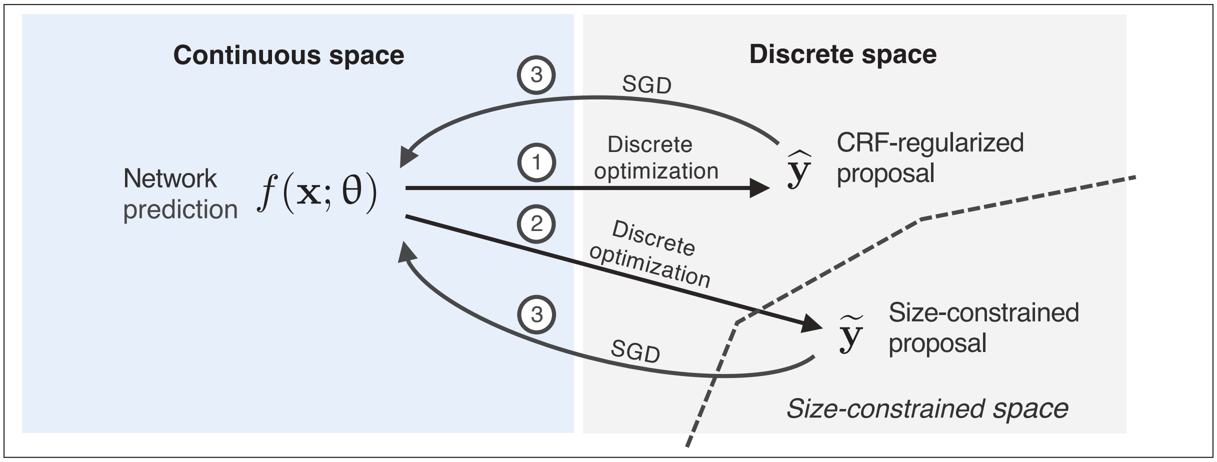


Figure 2.1 Diagram of the proposed discrete-continuous model for weakly-supervised segmentation. The learning process, based on the ADMM algorithm, alternates between the the following three steps: (1) the update of discrete CRF-regularized proposal $\widehat{\mathbf{y}}$, (2) the update of discrete size-constrained proposal $\widetilde{\mathbf{y}}$, and (3) the continuous update of a deep neural network f for segmentation, using gradient-descent. In the figure, $\mathbf{x}$ denotes an input image and θ are the parameters of the segmentation network.

summary of notation and frequently-used symbols is given in Table 2.1.

2.4.1 Problem formulation

Table 2.1 Notation and frequently used symbols.

$\mathcal{D}$	Dataset of examples $\{(\mathbf{x}^i, \mathbf{y}^i, \Omega^i)\}_{i=1}^N$.
$\mathbf{x}^i$	Image with n_i pixels.
$\mathbf{y}^i$	Ground-truth labels for image $\mathbf{x}^i$.
Ω^i	Subset of pixels in image $\mathbf{x}^i$ whose label is available during training.
$\mathbf{s}^i(\theta)$	Network's segmentation output for image $\mathbf{x}^i$, i.e. $\mathbf{s}^i(\theta) = f(\mathbf{x}^i;\theta)$.
$\mathcal{L}_{\text{lab}}, \mathcal{L}_{\text{reg}}$	Partial cross-entropy supervised loss and CRF-regularization (unsupervised) loss terms.
$\widehat{\mathbf{y}}^i, \widetilde{\mathbf{y}}^i$	CRF-regularized proposal and size-constrained proposal for image $\mathbf{x}^i$.
$\widehat{\mathbf{u}}^i, \widetilde{\mathbf{u}}^i$	Lagrange multipliers corresponding to the CRF-regularized proposal $\widehat{\mathbf{y}}^i$ and size-constrained proposals $\widetilde{\mathbf{y}}^i$.
$\widehat{\mu}^i, \widetilde{\mu}^i$	ADMM penalty parameters corresponding to the CRF-regularized proposal $\widehat{\mathbf{y}}^i$ and size-constrained proposals $\widetilde{\mathbf{y}}^i$.
$S_{\min}, S_{\max}$	Minimum and maximum size bounds on the segmentation foreground.

We focus on the following weakly-supervised segmentation problem where the labels of training images are only provided for a small subset of pixels. Given a training dataset $\mathcal{D} = \{(\mathbf{x}^i, \mathbf{y}^i, \Omega^i)\}_{i=1}^N$ where $\mathbf{x}^i$ is an image, $\mathbf{\Omega}^i$ is the subset of labeled pixels, and $\mathbf{y}^i$ the corresponding labels, we want to learn a segmentation network f parameterized by θ, such that $f(\mathbf{x}^i, \theta)$ gives the label probabilities at each pixel. For simplicity, we suppose a two-class segmentation problem and, using $\mathbf{s}^i(\theta) = f(\mathbf{x}^i; \theta)$ as short-hand notation, denote as $s_p^i \in [0, 1]$ the foreground probability predicted for a pixel p.

To learn parameters θ in this setting, we impose three requirements on the network output: 1) it should respect labeled pixels in Ω^i; 2) it should satisfy *discrete* segmentation constraints C_j, for $j = 1, \ldots, M$; 3) it should minimize a *discrete* regularization term $\mathcal{L}_{\text{reg}}$. Considering these requirements, we formulate the task as the following optimization problem:

$$\begin{aligned} \min_{\theta} \; \mathcal{L}(\theta) \; &= \; \sum_{i=1}^{N} \mathcal{L}_{\text{lab}} \left(\mathbf{s}^i(\theta), \mathbf{y}^i\right) \; + \; \lambda \mathcal{L}_{\text{reg}} \left(\mathbf{s}^i(\theta)\right) \\ \text{s.t.} \;\; C_j \left(\mathbf{s}^i(\theta)\right) \; &\leq \; 0, \;\; i = 1, \ldots, N, \;\; j = 1, \ldots, M, \end{aligned} \tag{2.1}$$

In this formulation, $\mathcal{L}_{\text{lab}}$ imposes the segmentation network to agree with provided pixel annotations, $\mathcal{L}_{\text{reg}}$ enforces the segmentation to be spatially regular, and λ is a hyper-parameter controlling the trade-off between label satisfaction and regularization.

The formulation in Eq. (2.1) is very general and covers a wide range of segmentation applications. For the purpose of this paper, however, we focus on a particular scenario similar to the one considered in Kervadec *et al.* (2019b). As in this previous work, we define $\mathcal{L}_{\text{lab}}$ as the cross-entropy loss over labeled pixels:

$$\mathcal{L}_{\text{lab}} \left(\mathbf{s}^i, \mathbf{y}^i\right) \; = \; - \sum_{p \in \Omega^i} y_p^i \log s_p^i \; + \; (1 - y_p^i) \log \left(1 - s_p^i\right). \tag{2.2}$$

Moreover, while any regularization prior can be used for $\mathcal{L}_{\text{reg}}$, we consider in this work a

Conditional Random Field (CRF) prior based on the weighted Potts model (Boykov *et al.*, 2001). This model encourages assigning the same label to pairs of pixels with similar features (i.e., pairwise consistency), and therefore provides robustness to noise (Zhang, Desrosiers & Zhang, 2018; Zhang & Desrosiers, 2018). Let $\tau(x) = 1(x \geq 0.5)$ be the function converting probability x to a binary value, the CRF-regularization prior is defined as $\mathcal{L}_{\mathrm{reg}}(\mathbf{s}^i) = \sum_{p,q} w^i_{p,q} \, |\tau(s^i_p) - \tau(s^i_q)|$, where $w^i_{p,q} = \exp\left(-\frac{1}{2\sigma^2}|x^i_p - x^i_q|^2\right)$ if pixels p and q are within a given neighborhood of each other, else $w^i_{p,q} = 0$. For neighbor pixels, the weight measures the similarity of pixel intensity or color. This regularization prior, which is frequently used for segmenting medical images, estimates the boundary length of the foreground region, hence minimizing it gives a segmentation with smooth contours. Last, as in Kervadec *et al.* (2019b) and Pathak *et al.* (2015), we impose lower and upper size bounds on the foreground: $S_{\min} \leq \sum_p \tau(s^i_p) \leq S_{\max}$. This type of constraint is also well-suited for medical image segmentation since the size of anatomical regions is typically limited by biology and size bounds can often be found in anatomical studies (e.g., see Medrano-Gracia *et al.* (2014) and Duane *et al.* (2017) for studies on cardiac structures).

As we will show in our experiments, the combination of boundary regularization and size constraints on the foreground is essential to achieve good performance in the weakly-supervised setting of this work. Since we only have partially-labeled images, size constraints are generally insufficient to uniquely define the foreground. Boundary regularization helps the network focus on regions that also respect edges in the image. However, since the contrast between the foreground and background varies largely from one image to another, using only this regularization with a fixed parameter λ can lead to under- or over-segmentation (e.g., the network predicting foreground or background at every pixel). Hence, adding size constraints to the regularization prior provides a way to control the effect of this prior on individual images.

In the next sections, we present two popular approaches, the penalty method and the ADMM algorithm, for solving constrained optimization problems like the one in Eq. (2.1). We then explain how the latter can be used to solve this problem efficiently with a mixed continuous-discrete formulation.

2.4.2 Penalty-based optimization

A popular method for constraining the output of a neural network is to model the constraint as a penalty term in the loss function (Kervadec *et al.*, 2019b; Pathak *et al.*, 2015). This additional loss term is typically a differentiable, convex function which equals zero for any output satisfying the constraint, and otherwise produces a positive value proportional to the degree of constraint violation. The main advantage of this approach is that it can be used directly within standard optimization techniques for training neural networks, for instance stochastic gradient descent (SGD), and typically leads to a smooth optimization (e.g., little oscillation due to the constraint). However, the penalty-based approach also suffers from important limitations which we illustrate by considering the simple problem of weakly-supervised segmentation with only an upper bound on the foreground size, i.e. $\sum_p s_p^i(\theta) \leq S_{\max}$. Using a squared penalty function, the problem can be expressed as

$$\mathcal{L}_{\text{pen}}(\theta) \;=\; \sum_{i=1}^{N} \mathcal{L}_{\text{lab}}\left(\mathbf{s}^i(\theta), \mathbf{y}^i\right) \;+\; \frac{\mu}{2}\Big(\max\Big\{0,\; \sum_{p} s_p^i(\theta) - S_{\max}\Big\}\Big)^2. \tag{2.3}$$

For the task at hand, the above formulation suffers from three important problems. First, it requires tuning penalty parameter μ, possibly for each image, otherwise the constraint may not be satisfied. Second, since the foreground size is estimated as the sum of probability values, instead of hard label assignments, it is very likely that the constraint will not be met once probability values are thresholded. For example, assigning a 50% foreground probability to all pixels in the image gives the same sum as giving a 100% probability to half the pixels. A possible strategy to enforce hard assignments in the network is to increase the temperature parameter of the softmax, however this leads to gradient saturation which freezes the solution in a local minima. The third problem can be understood by looking at the gradient of the loss with respect to network parameters, i.e.

$$\nabla_\theta \mathcal{L}_{\text{pen}}(\theta) \;=\; \sum_{i=1}^{N} \nabla_\theta \mathcal{L}_{\text{lab}}\left(\mathbf{s}^i(\theta), \mathbf{y}^i\right) \;+\; F^i \sum_{p} \nabla_\theta s_p^i(\theta), \tag{2.4}$$

with $F^i = \mu \cdot \max\left\{0,\, \sum_p s_p^i(\theta) - S_{\max}\right\}$. We see that, if the foreground size is greater than $S_{\max}$, the network simply scales down the gradient for each pixel by a constant factor F^i (note that the actual change in probability for a pixel is also proportional to its prediction uncertainty). This uniform scaling of gradient can result in a bad local minima if the shape to segment is complex (e.g., curved and narrow like the right ventricle) or the initial network output is poor.

2.4.3 Optimization with ADMM

Because the regularization prior and size constraints are discrete, the formulation in Eq. (2.1) cannot be optimized directly. To alleviate this problem, we propose and approach based on the alternating direction method of multipliers (ADMM) algorithm (Boyd *et al.*, 2011). ADMM is a variant of the augmented Lagrangian scheme which uses partial updates for the dual variables. It is often employed to solve problems in the form of $\min_{\mathbf{x}} f(\mathbf{x}) + g(\mathbf{x})$, where optimizing functions f and g together is hard, however minimizing each of them separately can be done more easily (e.g., to optimality and/or efficiently). The main idea is to reformulate the task as a constrained optimization problem $\min_{\mathbf{x},\mathbf{y}} f(\mathbf{x}) + g(\mathbf{y})$ subject to $\mathbf{x} = \mathbf{y}$, for which the objective function is separable in $\mathbf{x}$ and $\mathbf{y}$, and solve this new problem using an augmented Lagrangian method

$$\max_{\mathbf{u}} \min_{\mathbf{x},\mathbf{y}} \;\mathcal{L}_{\text{aug}}(\mathbf{x}, \mathbf{y}, \mathbf{u}) \;=\; f(\mathbf{x}) \;+\; g(\mathbf{y}) \;+\; \mathbf{u}^\top(\mathbf{x} - \mathbf{y}) \;+\; \frac{\mu}{2}\left\|\mathbf{x} - \mathbf{y}\right\|^2, \tag{2.5}$$

where $\mathbf{u}$ are the Lagrange multipliers and μ is the ADMM penalty parameter. An alternate definition, which we will employ in this paper, uses scaled multipliers (Boyd *et al.*, 2011):

$$\max_{\mathbf{u}} \min_{\mathbf{x},\mathbf{y}} \ \mathcal{L}_{\text{aug}}(\mathbf{x}, \mathbf{y}, \mathbf{u}) \ = \ f(\mathbf{x}) \ + \ g(\mathbf{y}) \ + \ \frac{\mu}{2} \left\| \mathbf{x} - \mathbf{y} + \mathbf{u} \right\|^2 . \tag{2.6}$$

Optimization of Eq. (2.6) is performed by solving iteratively with respect to each variable, while keeping others fixed, until convergence is reached:

$$\begin{aligned} \mathbf{x}^{t+1} &:= \ \text{argmin}_{\mathbf{x}} \ \mathcal{L}_{\text{aug}}(\mathbf{x}, \mathbf{y}^t, \mathbf{u}^t) \\ \mathbf{y}^{t+1} &:= \ \text{argmin}_{\mathbf{y}} \ \mathcal{L}_{\text{aug}}(\mathbf{x}^{t+1}, \mathbf{y}, \mathbf{u}^t) \\ \mathbf{u}^{t+1} &:= \ \mathbf{u}^t \ + \ (\mathbf{x}^{t+1} - \mathbf{y}^{t+1}) \end{aligned} \tag{2.7}$$

The ADMM algorithm can be used to efficiently solve certain types of constrained optimization problems. Thus, we can suppose that function g models a hard constraint, i.e. it returns 0 if its input satisfies the constraint, else it returns infinity. The main advantage of this formulation, compared to the penalty-based method in Eq. (2.3), is simplicity: we can minimize f in a unconstrained problem (update of $\mathbf{x}$), and finding a feasible solution amounts to solving a simple proximal problem (update of $\mathbf{y}$). As shown in the next section, this simplicity enables the use of efficient techniques in the discrete optimization setting. Another benefit of ADMM is that it dynamically adjusts the strength of the penalty, using Lagrange multipliers, to enforce the satisfaction of constraints. Hence, it is less sensitive to the choice of μ than the penalty method.

2.4.4 Discretely-constrained segmentation using ADMM

We now show how to use ADMM to solve the problem of Eq. (2.1) efficiently. We propose a mixed discrete-continuous formulation, where network parameters are optimized using standard gradient-based optimization, while the size-constrained segmentation problem is solved using specialized techniques for discrete optimization. Toward this goal, we decouple the label loss $\mathcal{L}_{\text{lab}}$, discrete regularization loss $\mathcal{L}_{\text{reg}}$ and discrete size constraint by introducing two binary segmentation vectors $\widehat{\mathbf{y}}^i$ and $\widetilde{\mathbf{y}}^i$ for each training image $\mathbf{x}^i$, and adding the following equality constraints: $\mathbf{s}^i(\theta) = \widehat{\mathbf{y}}^i$, $\mathbf{s}^i(\theta) = \widetilde{\mathbf{y}}^i$. These added vectors can be thought of as segmentation proposals, which are updated iteratively from the network's predictions, and help transfer discrete regularization and constraints to the training process. As we later present, this particular variable splitting strategy is chosen so that updating these proposals can be done efficiently and to optimality. Figure 4.1 summarizes the proposed strategy.

Using the ADMM formulation of Eq. (2.6), we then rewrite the segmentation problem as

$$\begin{aligned} \max_{\widehat{\mathbf{u}},\widetilde{\mathbf{u}}} \; \min_{\theta,\widehat{\mathbf{y}},\widetilde{\mathbf{y}}} \; \mathcal{L}_{\text{aug}}\left(\theta, \widehat{\mathbf{y}}, \widetilde{\mathbf{y}}, \widehat{\mathbf{u}}, \widetilde{\mathbf{u}}\right) \; = \; & \sum_{i=1}^{N} \mathcal{L}_{\text{lab}}\left(\mathbf{s}^i(\theta), \mathbf{y}^i\right) \; + \; \lambda \mathcal{L}_{\text{reg}}(\widehat{\mathbf{y}}^i) \\ & + \; \frac{\widehat{\mu}}{2} \left\|\mathbf{s}^i(\theta) - \widehat{\mathbf{y}}^i + \widehat{\mathbf{u}}^i\right\|_2^2 \; + \; \frac{\widetilde{\mu}}{2} \left\|\mathbf{s}^i(\theta) - \widetilde{\mathbf{y}}^i + \widetilde{\mathbf{u}}^i\right\|_2^2 \\ \text{s.t.} \;\; S_{\min} \; \leq \; \sum_p \widetilde{y}^i_p \; \leq \; & S_{\max}, \; \forall i \\ \widehat{\mathbf{y}}^i, \widetilde{\mathbf{y}}^i \; \in \; & \{0,1\}^{n_i}, \; \forall i. \end{aligned} \tag{2.8}$$

Here, n_i is the number of pixels in image $\mathbf{x}^i$. In the following subsections, we explain how each variable of this problem is updated (Lagrange multipliers are updated as per the standard ADMM method).

Network parameters

To update network parameters (i.e., convolution filter weights), we use a mini-batch gradient descent technique. Let $\mathcal{B} \subset \mathcal{D}$ be a batch of training samples, the gradient of the loss for batch $\mathcal{B}$ is given by

$$\sum_{i \in \mathcal{B}} \nabla_\theta \mathcal{L}_{\text{lab}}(\mathbf{s}^i, \mathbf{y}^i) \; + \; \sum_p \left(\widehat{\mu}(s^i_p - \widehat{y}^i_p + \widehat{u}^i_p) \; + \; \widetilde{\mu}(s^i_p - \widetilde{y}^i_p + \widetilde{u}^i_p)\right) \nabla_\theta s^i_p(\theta). \tag{2.9}$$

We see that, unlike the penalty method gradient in Eq. (2.4), the gradient at each pixel of image is scaled in a non-uniform manner based on the discrete CRF-regularized and size-constrained proposals. The parameters are then updated by taking a step opposite to the gradient, i.e. $\theta^{t+1} := \theta^{t} - \eta \nabla_{\theta} \mathcal{L}^{t}_{\text{aug}}$, where $\nabla_{\theta} \mathcal{L}_{\text{aug}}$ is the gradient defined in Eq. (2.9) and η is the learning rate. One should note that this gradient-based update does not guarantee an optimal solution for the network parameters, which is normally required for ADMM. Instead, this corresponds to a stochastic variant of ADMM that has been shown to convergence under weak conditions (Ouyang, He, Tran & Gray, 2013).

CRF-regularized proposal

Considering all other variables fixed, updating each CRF-regularized proposals $\widehat{\mathbf{y}}^{i}$ amounts to solving

$$\min_{\widehat{\mathbf{y}}^{i} \in \{0,1\}^{n_i}} \ \frac{1}{2} \left\| \widehat{\mathbf{y}}^{i} - (\mathbf{s}^{i}(\theta) + \widehat{\mathbf{u}}^{i}) \right\|_2^2 \ + \ ({}^{\lambda}/_{\widehat{\mu}})\, \mathcal{L}_{\text{reg}}(\widehat{\mathbf{y}}^{i}). \tag{2.10}$$

Using the property that $x^2 = x$ for a binary variable x, this problem can be expressed equivalently as a standard CRF energy minimization problem

$$\min_{\widehat{\mathbf{y}}^{i} \in \{0,1\}^{n_i}} \ \sum_{p} \widehat{a}^{i}_{p}\, \widehat{y}^{i}_{p} \ + \ ({}^{\lambda}/_{\widehat{\mu}})\, \mathcal{L}_{\text{reg}}(\widehat{\mathbf{y}}^{i}), \tag{2.11}$$

where $\widehat{a}^{i}_{p} = \frac{1}{2} - s^{i}_{p}(\theta) - \widehat{u}^{i}_{p}$ are unary potentials and $\mathcal{L}_{\text{reg}}$ is a pairwise (or higher-order) regularization prior. If $\mathcal{L}_{\text{reg}}$ is a sub-modular function, as the weighted Potts model used in this work, then the global optimum of this discrete optimization problem can be obtained in polynomial time with a max-flow algorithm (Boykov *et al.*, 2001). Note that this would not be the case if foreground size constraints were added to Eq. (2.11), which motivates the splitting strategy chosen for our method.

Size-constrained proposal

Likewise, we update each proposal $\widetilde{\mathbf{y}}^i$ by considering all other variables fixed and solving the following constrained discrete problem:

$$\min_{\widetilde{\mathbf{y}}^i \in \{0,1\}^{n_i}} \frac{1}{2} \left\| \widetilde{\mathbf{y}}^i - (\mathbf{s}^i(\theta) + \widetilde{\mathbf{u}}^i) \right\|_2^2, \quad \text{s.t. } S_{\min} \leq \sum_p \widetilde{y}_p^i \leq S_{\max}. \tag{2.12}$$

Using the same trick for binary variables as before, we then rewrite this problem as

$$\max_{\widetilde{\mathbf{y}}^i \in \{0,1\}^{n_i}} \sum_p \widetilde{a}_p^i \, \widetilde{y}_p^i, \quad \text{s.t. } S_{\min} \leq \sum_p \widetilde{y}_p^i \leq S_{\max}, \tag{2.13}$$

with $\widetilde{a}_p^i = s_p^i(\theta) - \widetilde{u}_p^i - \frac{1}{2}$. This discrete problem corresponds to a specific instance of the knapsack problem (Chu & Beasley, 1998), where each pixel p is an object with equal weight 1 and utility $\widetilde{a}_p^i$. The goal is to select between $S_{\min}$ and $S_{\max}$ objects such that their total utility is maximized. The optimal solution to this problem can be obtained via a simple ranking method where we set $\widetilde{y}_p^i = 1$ for the $S_{\min}$ pixels with highest utility and, if any, for the remaining pixels with highest *positive* utility until $S_{\max}$ is reached. Figure 2.2 illustrates this ranking procedure on a toy example.

2.4.5 Algorithm summary and complexity

The whole training process is summarized in Algorithm 3.1. Starting with zero-valued multipliers and proposals with equal foreground and background probabilities for each pixel, each training epoch involves the following steps. First, for $T_{\max}$ iterations, we update network parameters by randomly selecting a batch of 2D training images, computing the gradient for this batch and applying a descent step with this gradient. Next, we re-compute the 3D CRF-regularized

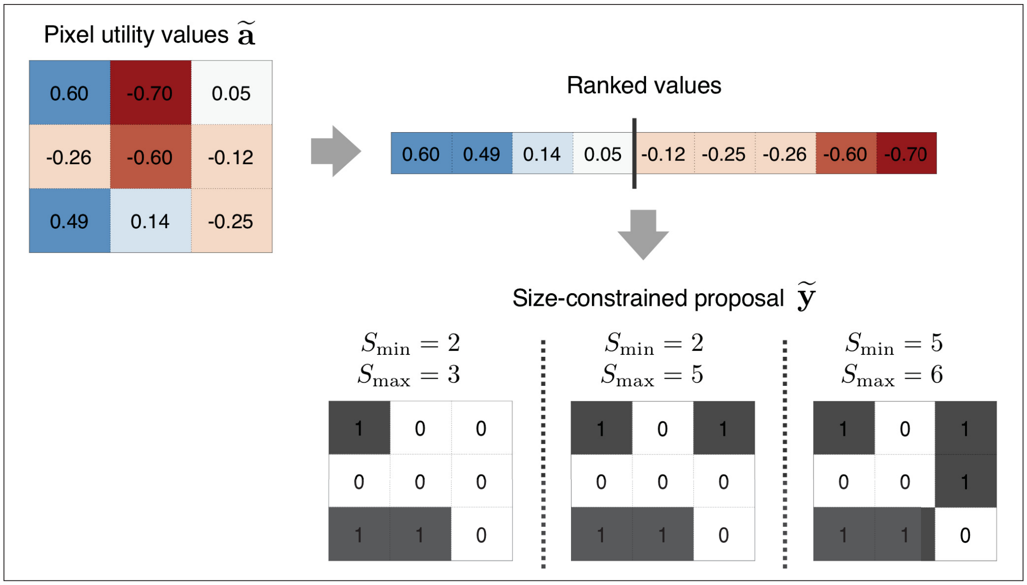


Figure 2.2 Illustration of the ranking method for updating the size-constrained proposal $\widetilde{\mathbf{y}}$. (1) Pixel utility values $\widetilde{\mathbf{a}}$ are first computed from the segmentation network and Lagrange multipliers: $\widetilde{a}_p = s_p(\theta) - \widetilde{u}_p - 1/2$. (2) Utility values are ranked by descending order, and (3) the discrete size-constrained proposal $\widetilde{\mathbf{y}}$ is obtained by selecting the $S_{\min}$ pixels with highest utility and, if any, the remaining pixels with highest *positive* utility, until $S_{\max}$ is reached. The resulting proposal is shown for three different size bounds $[S_{\min}, S_{\max}]$.

proposals ($\widehat{\mathbf{y}}^i$) and size-constrained proposals ($\widetilde{\mathbf{y}}^i$) using the modified network output. Last, we update Lagrange multipliers for both proposals and reduce the learning rate by a factor of decr_η. We note that the network doesn't need to be re-trained from scratch after each proposal update and that the only requirement for convergence is that the update of network parameters decreases the overall loss (Boyd *et al.*, 2011).

Since proposals are updated only once per epoch, our method yields negligible computational overhead compared to optimizing only the network with SGD. Moreover, each of these updates has low computational complexity. Thus, computing each $\widehat{\mathbf{y}}^i$ is done by solving a max-flow problem, which has $O(n^3)$ complexity where n is the number of image pixels/voxels. Likewise, updating each $\widetilde{\mathbf{y}}^i$ simply requires to sort pixels/voxels, the complexity of which is in $O(n \log n)$.

Algorithm 2.1 The proposed discretely-constrained segmentation

```
Input: Weakly-labeled images $\mathcal{D} = \{(\mathbf{x}^i, \mathbf{y}^i, \Omega^i)\}_{i=1}^{N}$;
Input: Size bounds $[S_{\min}, S_{\max}]$;
Output: Network parameters $\theta$;
/* Initialization */
Randomly initialize network parameters $\theta$;
Set $\widehat{y}^i_p, \widetilde{y}^i_p = 1/2$ and $\widehat{u}^i_p, \widetilde{u}^i_p = 0, \forall i, p$;
/* Main loop */
for epoch = $1, \ldots, E_{\max}$ do
    /* Network parameters update */
    for iter = $1, \ldots, T_{\max}$ do
        Randomly select batch $\mathcal{B} \subset \mathcal{D}$;
        Apply batch gradient step as in Section 2.4.4;
    end for
    /* Discrete proposals update */
    Update CRF-regularized proposal $\widehat{\mathbf{y}}^i$ as in Section 2.4.4;
    Update size-constrained proposal $\widetilde{\mathbf{y}}^i$ as in Section 2.4.4;
    /* Multipliers update */
    $\widehat{\mathbf{u}}^i := \widehat{\mathbf{u}}^i + (\mathbf{s}^i(\theta) - \widehat{\mathbf{y}}^i), \forall i$;
    $\widetilde{\mathbf{u}}^i := \widetilde{\mathbf{u}}^i + (\mathbf{s}^i(\theta) - \widetilde{\mathbf{y}}^i), \forall i$;
    Decrease learning rate: $\eta := \mathrm{decr}_\eta \times \eta$, with $\mathrm{decr}_\eta \in [0, 1]$;
end for
return $\theta$ ;
```

2.5 Experiments

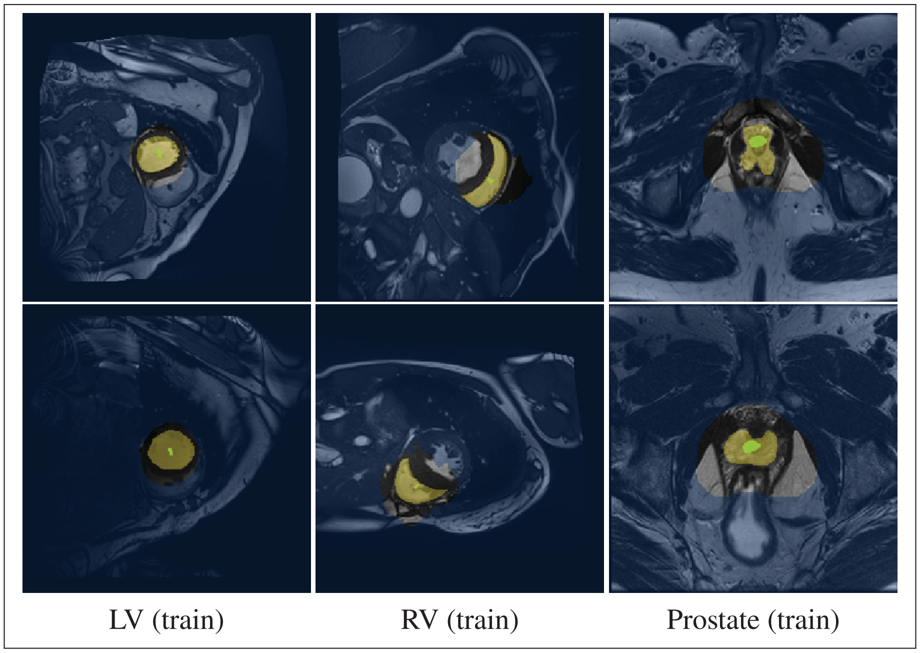
2.5.1 Datasets and evaluation protocol

We evaluate the proposed method on three different medical imaging segmentation tasks: left-ventricular (LV) and right-ventricular (RV) endocardium segmentation in cine magnetic resonance imaging (MRI), and prostate segmentation in T2-MRI.

LV and RV segmentation: This medical image set is provided by the Automated Cardiac Diagnosis Challenge (ACDC) (Bernard *et al.*, 2018) and focuses on the segmentation of three cardiac structures, i.e. left ventricular endocardium and epicardium, and right ventricular endocardium. It consists of 100 cine MRI exams covering normal cases and subjects with well-defined defined pathologies: dilated cardiomyopathy, hypertrophic cardiomyopathy, myocardial infarction with altered left ventricular ejection fraction and abnormal right ventricle. Each exam contains acquisitions at the diastolic and systolic phases. The spatial resolution goes from 0.83 to 1.75 mm^2/pixel, with a thickness of 5-8 mm and an inter-slice gap of 5 mm, covering the LV from the base to the apex. In our experiments, 80 exams were employed for training and the remaining 20 for validation.

Prostate segmentation: For the third task, we used the dataset made available at the MICCAI 2012 prostate MR segmentation challenge (PROMISE) (Litjens *et al.*, 2014). This dataset contains multi-centric transversal T2-weighted MR images from 50 subjects acquired with multiple MRI vendors and different scanning protocols, which are representative of typical MR images acquired in a clinical setting. The images resolution ranges from $15 \times 256 \times 256$ to $54 \times 512 \times 512$ voxels with a spacing ranging from $2 \times 0.27 \times 0.27$ to $4 \times 0.75 \times 0.75$ mm^3. In this case, we employed 40 patients for training and 10 for validation during the experiments.

We generate weak annotations based on the scenario where an annotator identifies for each volume a single position (e.g., mouse click) near the centroid of the foreground region, and a global anatomical atlas is then used to obtain foreground and background seeds from this position. The following procedure is employed to derive annotations under this scenario. Considering training images only, we compute the centroid of each 3D ground truth mask and use it as the origin of a global coordinate system, i.e. position $(0,0,0)$ in each volume corresponds to the foreground centroid in that volume. For each position in this new coordinate system, we then measure the fraction of training volumes for which the position is in the foreground region. Next, we define a global foreground mask containing all positions having a foreground probability of 100%, and a global background mask containing all those with 0% probability. Finally,



LV (train) RV (train) Prostate (train)

Figure 2.3 Examples of images, ground truth and annotations for the three test datasets. Green pixels correspond to foreground seeds, light yellow pixels to the ground-truth foreground region, and blue-shaded pixels to the background mask. Note that we only show a single slice of a 3D volume, and many slices (located further away from the foreground center) actually contain no foreground seed.

annotations are obtained for each training volume by aligning the global masks to the volume's foreground centroid, and using the aligned masks to label voxels in the volume accordingly.

We note that this weakly-supervised setting is more challenging than the one used in Kervadec *et al.* (2019b), where pixels were annotated for each slice of the volume containing the foreground region. In our case, slices away from the foreground center often have no annotation.

Figure 2.3 shows examples of annotations and ground truth for the three test datasets. It can be seen that the proposed procedure for generating weak annotations leads to diverse segmentation tasks with various levels of difficulty. For left ventricle segmentation (LV), foreground seeds represent on average 14.29% of the foreground region and 0.12% of the whole volume. In comparison, they only represent 2.29% of the foreground and 0.02% of the volume for right ventricle segmentation (RV). Moreover, for non-circular regions like RV, we see that foreground seeds can be close to the region boundary, making the segmentation more challenging.

We test three different settings for the proposed method, using 1) only CRF-regularized proposals, 2) only size-constrained proposals, and 3) both CRF-regularized and size-constrained proposals. For the two settings using size-constrained proposals, we define size bounds of different tightness by adding or substracting a relative percentage $\epsilon \in \{0\%, 10\%, 20\%, 40\%\}$ of the real foreground size: $[S_{\min}, S_{\max}] = [(1-\epsilon) \quad \times S_{\text{real}}, (1+\epsilon)\times S_{\text{real}}]$. Hence, $\epsilon = 0$ imposes the predicted segmentation to have the same size as the ground truth.

To get an upper bound on performance, we trained our segmentation model using fully-annotated images and call this baseline Fully-supervised in the results. We also compared our method against three recent techniques for weakly-supervised segmentation. The first one is the Penalty approach (Kervadec *et al.*, 2019b) presented in Section 2.4.2, that estimates the foreground size by summing the probabilities predicted for this region over the whole image, and then applies a squared-error loss on the difference between the estimated size and the lower or upper size bounds. The second is the constrained CNN approach of Pathak et al. (Pathak *et al.*, 2015), which computes a latent distribution of foreground probabilities by adding a Lagrange multiplier to CNN outputs inside the softmax. Note that this approach was originally proposed for 2D segmentation, and we adapted it to handle size constraints on 3D regions. The third approach is the well-known DeepCut algorithm (Rajchl *et al.*, 2017), which alternates between updating the network with SGD and regularizing the network's predicted segmentation using a dense CRF. For all three approaches, we employed the same network architecture, seeds and partial cross-entropy loss as for our method. The hyper-parameters of these approaches were tuned using grid search.

The performance of segmentation models is evaluated using the 3D Dice similarity coefficient

(DSC). This well-known metric measures the overlap between the predicted segmentation S and ground truth mask G, as

$$\mathrm{DSC}(S, G) = \frac{2|S \cap G|}{|S|+|G|}. \tag{2.14}$$

2.5.2 Implementation details

For all experiments on the ACDC dataset, we followed Kervadec *et al.* (2019b) and used ENet (Paszke, Chaurasia, Kim & Culurciello, 2016) as our segmentation architecture. This lightweight network, which has been used in urban segmentation and medical imaging, gives a good trade-off between accuracy and inference speed. The ENet architecture is mainly composed of Bottleneck modules which process the information in two separate branches, one performing max pooling followed by padding and the other one applying a sequence of 1×1, regular, dilated or full convolution, 1×1 convolution and spatial Dropout. The output of the two branches is then merged with an element-wise addition, as in ResNet blocks. More information can be found in (Paszke *et al.*, 2016).

We trained the network from scratch with SGD using the Adam optimizer and a batch size of 8. As in (Kervadec *et al.*, 2019b), 3D volumes are segmented slice-by-slice using images of size 256×256 as input to the network, without data augmentation. The initial learning rate was set to 5×10^{-4} and decreased by a factor of 4 every 50 epochs for a total of 250 training epochs. For CRF regularization, we used $\lambda = 100$ for all experiments and selected a different σ to account for contrast differences between the left-ventricle (LV) and right-ventricle (RV): $\sigma = 10^{-4}$ for LV and $\sigma = 10^{-5}$ for RV.

For the PROMISE dataset, we employed the U-net architecture (Ronneberger *et al.*, 2015) with 15 layers, batch normalization and ReLU activation. This architecture is one of the most popular models for segmentation, especially for tasks related to bio-medical imaging. The U-net architecture is made of a contracting path, with a repeated application of two 3×3 convolutions with ReLU and 2×2 max pooling with stride 2, and an expansive path which applies a sequence of upsampling with 2×2 convolution, concatenation with the corresponding feature map of the contracting path, and two 3×3 convolutions with ReLU. A full description of this architecture can be found in (Ronneberger *et al.*, 2015). While batch normalization was not used in the original work, our experiments showed it to significantly improve training speed and stability.

We trained the network from scratch using the Adam optimizer with a weight decay of 10^{-4} and a batch size of 4. The initial learning rate was set to 10^{-4} and decreased by a factor of 5 every 50 epoch for a total of 250 epochs. Since this segmentation task is more challenging, to have a good performance for the fully-supervised baseline, we performed the following data augmentation procedure: images resized to 256×256 pixels, followed by random rotation, random crop, and random flip. CRF regularization parameters were set to $\lambda = 1000$ and $\sigma = 10^{-6}$.

The same value was employed for the ADMM penalty parameters of our CRF + size method (i.e., $\widehat{\mu} = \widetilde{\mu}$). This value, as well as the value of the penalty method parameter μ, were selected for each segmentation task using grid search. Unless specified otherwise, we report for all methods the results obtained using the best value found for these parameters.

2.6 Results

In this section, we evaluate the performance of the proposed strategy for weakly-supervised segmentation. Moreover, we assess the benefit of using CRF-regularized and size-constrained proposals for training the network, and measure segmentation accuracy for different size bounds. Last, we evaluate the impact of the ADMM penalty parameter on results.

Table 2.2 Mean 3D Dice of tested methods on validation images for left ventricle (LV) and right ventricle (RV) segmentation of the ACDC dataset and prostate segmentation of the PROMISE dataset. For the Penalty method, the approach by Pathak et al. (Pathak *et al.*, 2015) and our method using size proposals (size only or CRF + size), we report accuracy for different foreground size bounds defined by ϵ. Values in the table are the average over three runs with different random initialization of network parameters (standard deviation in parentheses).

	Method	**LV**	**RV**	**Prostate**
	Fully-supervised	0.927 (0.012)	0.870 (0.009)	0.873 (0.011)
Without size bound	DeepCut (Rajchl *et al.*, 2017)	0.789 (0.018)	0.557 (0.016)	0.724 (0.031)
	Ours (CRF only)	0.862 (0.017)	0.677 (0.014)	0.769 (0.019)
Size bound: $\epsilon = 0\%$	Penalty (Kervadec *et al.*, 2019b)	0.813 (0.025)	0.529 (0.032)	0.588 (0.027)
	Pathak et al. (Pathak *et al.*, 2015)	0.848 (0.013)	0.609 (0.025)	0.615 (0.029)
	Ours (size only)	0.871 (0.021)	0.598 (0.028)	0.760 (0.036)
	Ours (CRF + size)	**0.901 (0.009)**	**0.730 (0.029)**	**0.807 (0.013)**
Size bound: $\epsilon = 10\%$	Penalty (Kervadec *et al.*, 2019b)	0.840 (0.036)	0.570 (0.028)	0.621 (0.029)
	Pathak et al. (Pathak *et al.*, 2015)	0.840 (0.040)	0.598 (0.038)	0.667 (0.027)
	Ours (size only)	0.844 (0.038)	0.617 (0.033)	0.766 (0.040)
	Ours (CRF + size)	**0.884 (0.017)**	**0.719 (0.032)**	**0.795 (0.025)**
Size bound: $\epsilon = 20\%$	Penalty (Kervadec *et al.*, 2019b)	0.833 (0.030)	0.498 (0.073)	0.583 (0.064)
	Pathak et al. (Pathak *et al.*, 2015)	0.848 (0.024)	0.599 (0.056)	0.635 (0.026)
	Ours (size only)	0.845 (0.026)	0.600 (0.035)	0.760 (0.032)
	Ours (CRF + size)	**0.872 (0.015)**	**0.734 (0.010)**	**0.809 (0.024)**
Size bound: $\epsilon = 40\%$	Penalty (Kervadec *et al.*, 2019b)	0.826 (0.021)	0.582 (0.033)	0.515 (0.067)
	Pathak et al. (Pathak *et al.*, 2015)	0.853 (0.012)	0.582 (0.020)	0.644 (0.017)
	Ours (size only)	0.827 (0.034)	0.570 (0.031)	0.758 (0.028)
	Ours (CRF + size)	**0.879 (0.019)**	**0.691 (0.024)**	**0.771 (0.029)**

2.6.1 Segmentation performance

Table 2.2 reports the results of our discretely-constrained method under three settings, i.e. using only CRF-regularized proposals (CRF only), only size-constrained proposals (size only), or both types of proposals (CRF + size), comparing them to the fully-supervised baseline, the penalty method in (Kervadec *et al.*, 2019b), the DeepCut algorithm (Rajchl *et al.*, 2017), and the constrained CNN approach of Pathak et al. (Pathak *et al.*, 2015).

Overall, the proposed method achieves a higher mean 3D Dice than the penalty approach for the

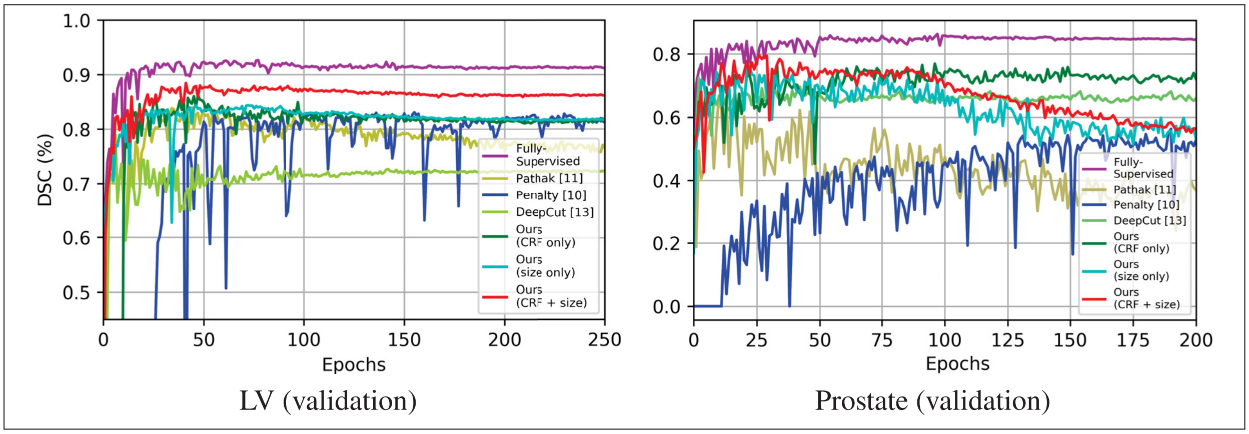


Figure 2.4 3D Dice of tested methods on validation images of the ACDC (LV) and PROMISE prostate datasets at each iteration of training. Note: foreground size bounds corresponding to $\epsilon = 10\%$ are employed for the penalty method and our method using size proposals (size only or with CRF).

same size bounds. This improvement ranges from 4-5% in the case of LV to nearly 20% for the RV and prostate segmentation tasks, and is consistent across nearly all tasks and configurations. We then examine the impact of using image-specific size constraints with different ϵ, when no CRF regularization is used (i.e., size only). As bounds on the size become less restrictive (i.e., larger ϵ), the segmentation performance of the proposed method degrades, resulting in a decrease of 3-4% with respect to the setting with $\epsilon = 0\%$. This indicates that an accurate size estimation can help improve segmentation in a weakly-supervised setting.

We also observe that employing a regularization prior jointly with size constraints (i.e., CRF + size) leads to a better performance, with DSC improvements of 3% to 15% compared to using only size proposals. Improvements are particularly important for the RV segmentation task which presents a more complex shape (i.e., curved and narrow) than LV or prostate, suggesting that imposing size constraints alone is not sufficient to get a satisfactory segmentation on certain structures. By including the CRF regularization, the segmentation is attracted towards the target contours, resulting in higher 3D Dice. The benefits of combining both priors can also be appreciated when comparing against the model with only CRF regularization. A drawback of

CRF regularization is its sensitivity to parameter λ, which should normally be tuned per image to avoid under- or over-segmentation. Adding a size prior adds robustness to this parameter, even when size constraints are less restrictive. Particularly, results show an accuracy boost of 3-5% over CRF regularization alone when employing tight bounds (ϵ = 0%), and a gain of 1-2% when bounds are loosest (ϵ = 40%).

Compared to DeepCut, our method without size bounds (CRF only) yields DSC improvements of 4%-12% over the three segmentation tasks. Likewise, significant improvements in performance are observed when comparing our method with only size bounds (size only) against the constrained CNN approach of Pathak et al., for all size bound ϵ values. Since both these comparison baselines use a continuous optimization model, this illustrates the advantage of employing a discrete formulation for regularized and size-constrained segmentation.

Figure 2.4 depicts the 3D Dice of tested methods measured at each training epoch on validation images of the LV and prostate segmentation tasks. We include this figure to visualize the convergence of methods and their stability over training. Compared to the penalty strategy, the proposed method converges faster and generally shows a more stable behaviour. This underlines the importance of proposals to guide the network in the beginning of training, when images are partly labeled. Note that the highest validation accuracy of our method is obtained early in the training (epoch 50), showing the benefit of a discrete formulation where size constraints are satisfied at each iteration instead of incrementally.

Figure 2.5 illustrates the evolution of the CRF-regularized proposal, size-constrained proposal and network prediction at different epochs for a training image. We observe that the CRF proposal, although not perfect, plays an important role in the beginning. As training progresses, the size proposal then helps correct errors of the CRF by constraining the foreground size. At convergence, the network prediction and proposals are nearly identical. As mentioned above, the best segmentation is obtained early in training (see epoch 10 in the figure).

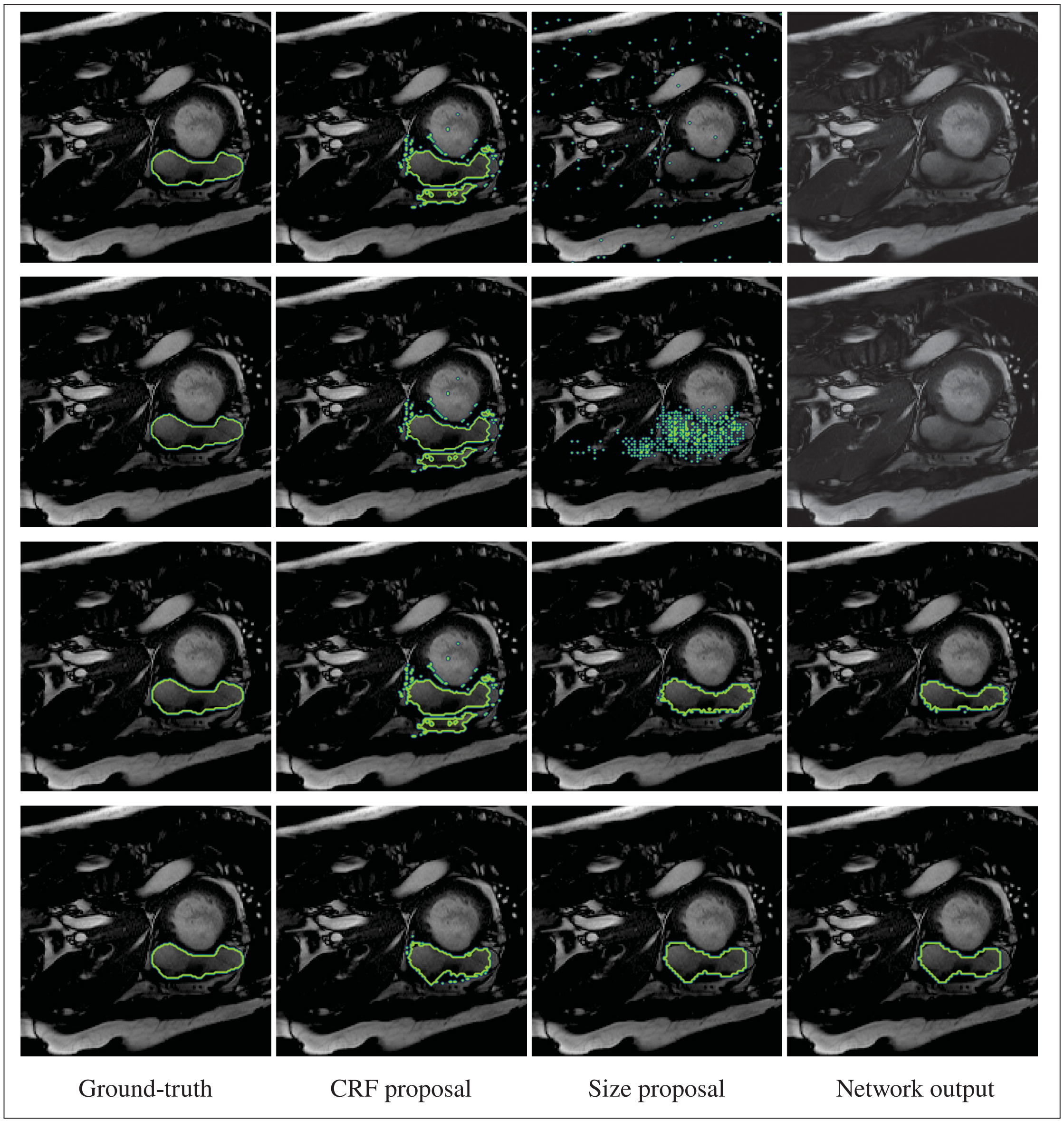
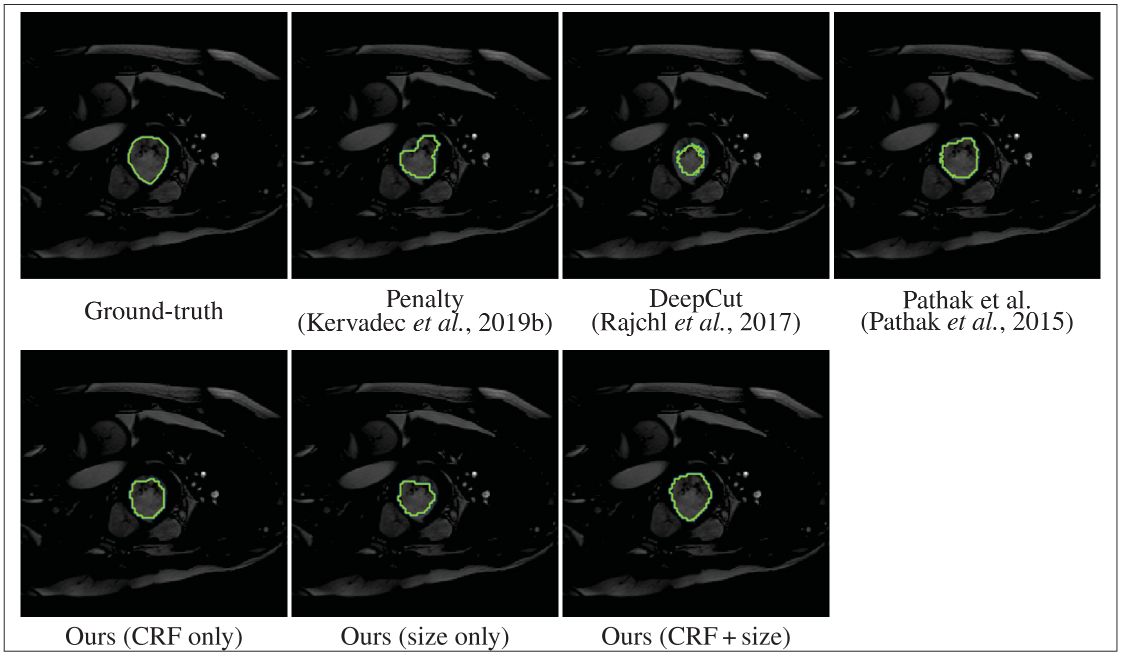


Figure 2.5 Evolution of the CRF-regularized proposal $(\widehat{\mathbf{y}})$, size-constrained proposal $(\widetilde{\mathbf{y}})$ and network output at different training stages of our CRF + size method for $\epsilon = 0$. From top to bottom: epoch= 0, epoch= 2, epoch= 10 and epoch= 250.

2.6.2 Qualitative evaluation

Visual results of tested methods for the three segmentation tasks are depicted in Figures 2.6, 2.7 and 2.8. We observe that incorporating only size constraints during training might be insufficient to segment complex structures in a weakly-supervised scenario, regardless of the method used.



Ground-truth

Penalty (Kervadec *et al.*, 2019b)

DeepCut (Rajchl *et al.*, 2017)

Pathak et al. (Pathak *et al.*, 2015)

Ours (CRF only)

Ours (size only)

Ours (CRF + size)

Figure 2.6 Visual comparison of tested methods on a validation image of the ACDC left ventricle (LV) dataset for size bounds corresponding to $\epsilon = 10\%$.

For example, in the RV segmentation task (middle two rows), both the penalty approach and our method with only size proposals produce segmentation masks which are not well aligned with the image target boundaries, although their size are similar to the ground truth. Nevertheless, imposing discrete size constraints helps generate contours whose sizes are closer to the real target size, as shown in the last row of the figure. Inspecting the contours obtained by our model using only CRF regularization, we observe that they better match the target boundaries. However, the well-known shrinking bias problem of CRFs makes this model under-segment regions in some cases (e.g., see first two rows). The proposed model can overcome this issue by integrating both size and CRF regularization priors.

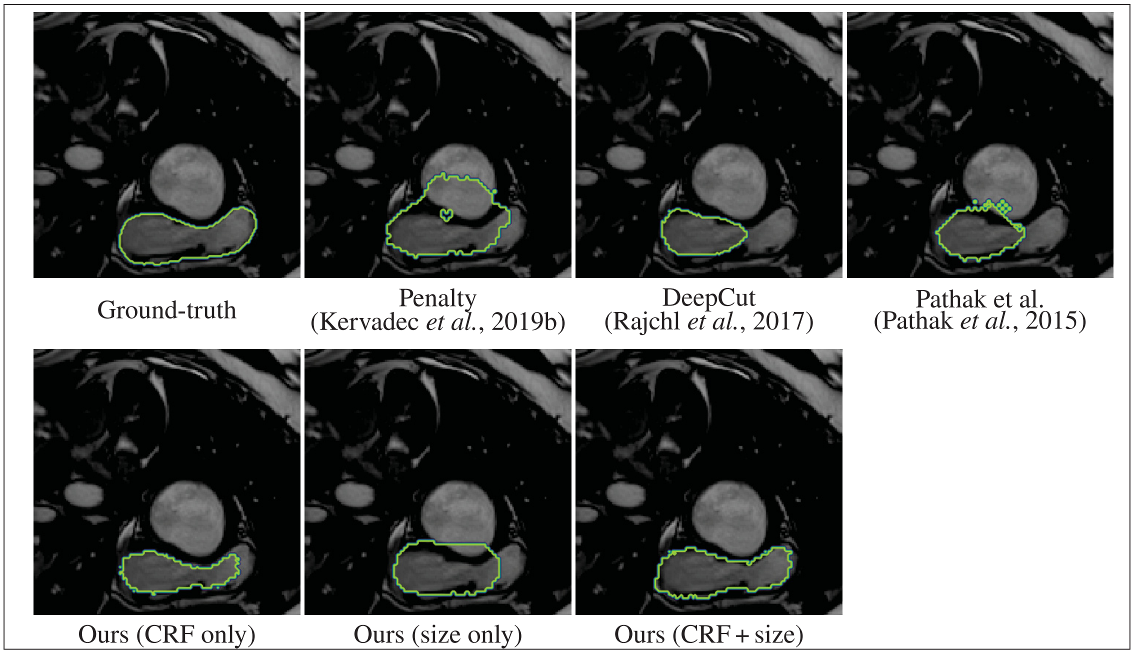
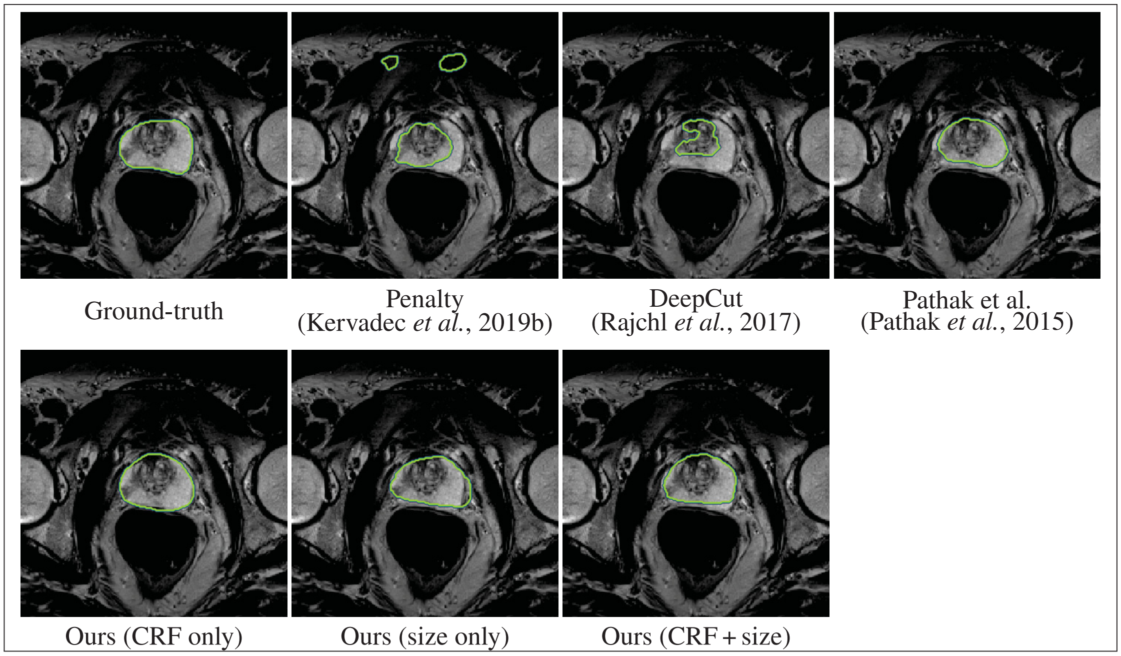


Figure 2.7 Visual comparison of tested methods on a validation image of the ACDC right ventricle (RV) dataset for size bounds corresponding to $\epsilon = 10\%$.

2.6.3 Constraint satisfaction

In Fig. 2.9, we evaluate our method's ability to satisfy imposed constraints. We consider size bounds corresponding to $\epsilon = 10\%$ and compute the ratio between the size of the predicted segmentation and the real foreground size for the different model settings. We first observe that the model with only CRF regularization (green curve) does not control the size of the foreground, as its predictions are pushed towards the visible boundaries in the image regardless of the target size. While the penalty approach (blue curve) converges to a mean ratio within size bounds, its behaviour remains unstable and generates segmentations whose size lies outside the imposed bounds, both during training and validation. On the other hand, if the size constraints are imposed in the discrete domain, the ratio between predicted and real sizes is also located within the bounds, but follows a more stable regime. This pattern is also observed in the configuration with both size and CRF priors (red curve). However, this configuration converges faster to predictions satisfying constraints, particularly in the validation set.


Ground-truth

Penalty (Kervadec *et al.*, 2019b)

DeepCut (Rajchl *et al.*, 2017)

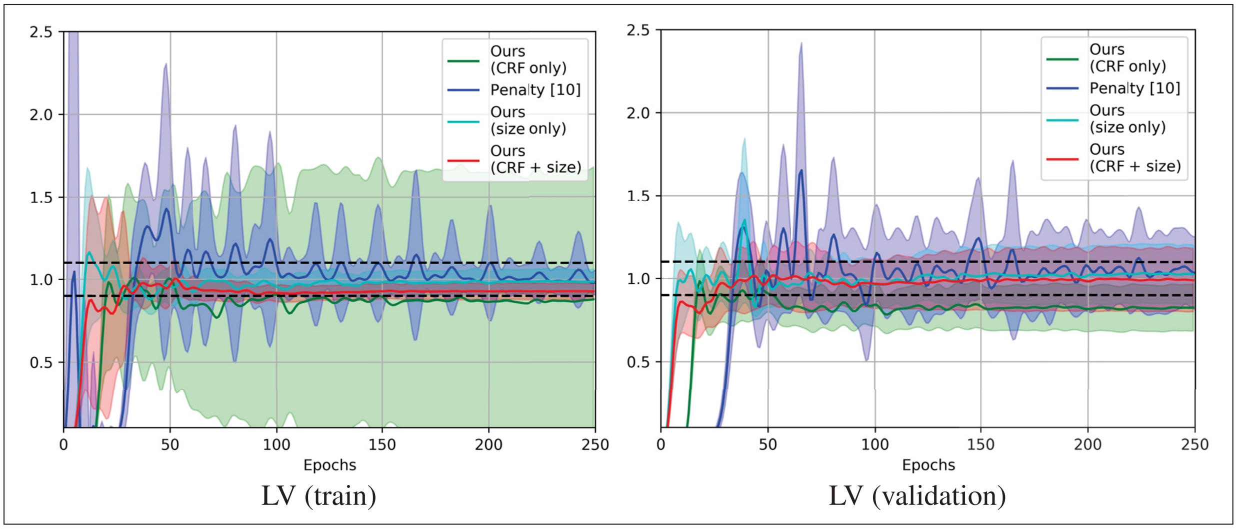
Pathak et al. (Pathak *et al.*, 2015)

Ours (CRF only)

Ours (size only)

Ours (CRF + size)

Figure 2.8 Visual comparison of tested methods on a validation image of the PROMISE prostate dataset for size bounds corresponding to $\epsilon = 10\%$.



LV (train)

LV (validation)

Figure 2.9 Ratio between predicted and real foreground size during training, while using size bounds corresponding to $\epsilon = 10\%$ (dashed lines). Solid lines correspond mean value and light-colored intervals to standard deviation.

Table 2.3 Mean 3D Dice obtained with different values for the ADMM penalty parameters $\widehat{\mu}$ and $\widetilde{\mu}$. For our CRF + size method, we used the same value for both parameter (i.e., $\widehat{\mu} = \widetilde{\mu}$). Tight bounds on foreground were employed for these results (i.e., $\epsilon = 0$). Values in the table are the average over three runs with different random initialization of network parameters (standard deviation in parentheses).

Method	$\widehat{\mu} = \widetilde{\mu}$	**LV**	**RV**	**Prostate**
CRF only	0.01	0.798 (0.019)	0.541 (0.020)	0.703 (0.025)
	0.1	0.825 (0.020)	0.594 (0.017)	0.744 (0.027)
	1	**0.862 (0.017)**	**0.677 (0.014)**	**0.769 (0.019)**
	10	0.828 (0.010)	0.675 (0.014)	0.768 (0.012)
Size only ($\epsilon = 0$)	0.01	0.797 (0.038)	0.548 (0.034)	0.695 (0.014)
	0.1	0.814 (0.024)	**0.598 (0.028)**	0.741 (0.021)
	1	**0.871 (0.021)**	0.576 (0.028)	**0.760 (0.036)**
	10	0.724 (0.030)	0.341 (0.073)	0.645 (0.049)
CRF + size ($\epsilon = 0$)	0.01	0.783 (0.020)	0.549 (0.019)	0.744 (0.010)
	0.1	0.835 (0.023)	0.604 (0.022)	0.740 (0.017)
	1	0.878 (0.013)	0.722 (0.024)	0.756 (0.009)
	10	**0.901 (0.009)**	**0.730 (0.029)**	**0.807 (0.013)**

2.6.4 Impact of the ADMM penalty parameter

As last experiment, we measure the sensitivity of our method to the ADMM penalty parameters $\widehat{\mu}$ and $\widetilde{\mu}$. As described in Section 2.4.4, these parameters control the trade-off between the constraint satisfaction loss (i.e., quadratic penalty term) and the supervised loss (i.e., partial cross-entropy). For our CRF + size method, the same value was used for both penalty parameters. Table 2.3 gives the mean 3D Dice obtained by the three settings of our method for different values of the penalty parameter. While the optimal value varies from one setting to the other, i.e. the best value for CRF only and size only is 1 while it is 10 for CRF + size, the best value for a given setting seems stable across different segmentation tasks.

2.7 Discussion and conclusion

We presented a novel method for training a CNN with discrete constraints and regularization priors. This method uses ADMM to split the continuous optimization of network parameters with SGD from the computation of discrete segmentation proposals. By incorporating both constraints and regularization priors, the network can be trained efficiently with weak annotations. We applied the proposed method to the segmentation of cardiac structures and prostates from MRI data, using partially-labeled images and bounds on the foreground size. Experiments show our method to provide significant improvements compared to the penalty approach of (Kervadec *et al.*, 2019b), in terms of segmentation accuracy, constraint satisfaction and convergence speed.

One of main advantages of our method compared to the optimization of constraints directly within a gradient-based framework is its ability to perform large steps in the solution space, with guaranteed optimality. However, this can also lead to instability during training (e.g., pixels in the discrete proposals oscillating between the foreground and background classes) if the penalty term is too strong. As described in (Boyd *et al.*, 2011), a possible way to alleviate this problem is to update the ADMM penalty parameter dynamically during training, for example starting with a smaller value and increasing it over epochs.

In the continuous optimization approach of Pathak et al. (Pathak *et al.*, 2015), size constraints are enforced by adding a Lagrange multiplier to network outputs within the softmax. A drawback of this approach is that it saturates the softmax, which causes the gradient to vanish and thus the learning to stop. The discrete formulation of our method avoids this problem by separating the Lagrange penalty term from the network output. Furthermore, Pathak et al.'s approach requires several dual update steps to satisfy size constraints for each image, and finding the optimal learning rate for each of these steps can be challenging: a too small rate leads to a slow satisfaction of constraints, while a too large one causes instability. In contrast, our discrete update of size-constrained proposals described in Section 2.4.4 guarantees the satisfaction of size constraints at each iteration.

Another important advantage of our method is the decoupling of the supervised loss from priors, using proposals. This allows training the network with batches containing 2D images sampled from different 3D volumes, which cannot be done in direct optimization approaches like the penalty method. It also enables incorporating more complex constraints and regularization priors in the learning process, a promising research direction that we will investigate in future work.

CHAPTER 3

DEEP CO-TRAINING FOR SEMI-SUPERVISED IMAGE SEGMENTATION

Jizong Peng[1] , Guillermo Estrada[2] , Marco Pedersoli[3] , Christian Desrosiers[1]

[1] Department of Software and IT Engineering, École de Technologie Supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3
[2] PUC-Rio, 225 Marquês de São Vicente Street, Rio de Janeiro, Brazil
[3] Department of Automated Production, École de Technologie Supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3

Article published in Journal *«Pattern Recognition»* in February 2020.

3.1 Presentation

This chapter presents the article “Deep Co-training for Semi-supervised Image Segmentation" by Peng, Estrada, Pedersoli, and Desrosiers accepted to the journal *Pattern Recognition* for publication on 10 February 2020. The objective of this research is to propose a new ensemble method for deep segmentation models. This scheme allows different models to exchange information among each other using the proposed consistency and adversarial losses, similar to co-training. We demonstrate the potential of our method on three challenging image segmentation problems, and illustrate its ability to share information between simultaneously trained models, while preserving their diversity.

3.2 Introduction

Semantic segmentation (Noh, Hong & Han, 2015) is a fundamental problem in computer vision, which requires assigning the proper category label to each pixel of a given image. It plays a key role in applications of various domains, including image retrieval, autonomous driving, video surveillance, remote sensing, robotics and biomedical imaging. This task is particularly important for medical image analysis, where it serves as a necessary pre-processing step for the assessment and treatment planning of various medical conditions (Litjens *et al.*, 2017).

In recent years, supervised approaches, in particular those based on deep learning, have shown tremendous potential for automated image segmentation. In such approaches, parametric models like fully-convolutional neural networks (F-CNNs) (Long *et al.*, 2015) are trained with a large set of annotated images by minimizing some loss function like cross-entropy or Dice loss (Milletari, Navab & Ahmadi, 2016). In many cases, however, obtaining sufficient data for training can be challenging, and manually annotating images can be a time consuming task (Kolesnikov & Lampert, 2016). This problem is even more significant in medical imaging applications, where images are typically 3D volumes (e.g., MRI or CT scans), the regions to delineate have low contrast, and annotations must be made by highly-trained experts. For challenging problems like infant brain segmentation, obtaining reliable annotations for a single subject may take a radiologist up to a week[1] (Wang *et al.*, 2019a).

To alleviate the need for fully-annotated data, numerous works have focused on developing weakly-supervised methods for segmentation. In such methods, easier to obtain annotations like image-level tags (Pinheiro & Collobert, 2015b; Papandreou, Chen, Murphy & Yuille, 2015; Kervadec *et al.*, 2019b; Pathak *et al.*, 2015), bounding boxes (Dai *et al.*, 2015; Rajchl *et al.*, 2017) or scribbles (Lin *et al.*, 2016) are used for training segmentation models, instead of whole-image pixel labels. Multiple instance learning (MIL) (Vezhnevets & Buhmann, 2010) is a popular technique for dealing with image tags, where images are considered as bags of pixels / superpixels (i.e., instances) and positive examples for a given object of interest (i.e., tag) are images for which at least one pixel / superpixel corresponds to that object. MIL methods for segmentation typically rely on objectness (Bearman, Russakovsky, Ferrari & Fei-Fei, 2016; Wei *et al.*, 2016; Pinheiro & Collobert, 2015a; Qi, Liu, Shi, Zhao & Jia, 2016; Saleh *et al.*,

[1] See http://iseg2017.web.unc.edu/reference/

2016; Shimoda & Yanai, 2016), class-specific saliency and activation maps (Hou *et al.*, 2017; Liu & Han, 2016; Selvaraju *et al.*, 2017; Zhou *et al.*, 2016), or image-level constraints (Kervadec *et al.*, 2019b; Pathak *et al.*, 2015) to obtain a prior on the presence or location of objects in the image.

In various scenarios, weakly-supervised learning methods for segmentation may not be suitable. For instance, adding bounding boxes or point annotations can still be time-costly for 3D scans, which may contain over 100 separate images (i.e., 2D slices). Likewise, image-level tags may not be useful in segmentation tasks where one must separate a single region of interest (i.e., foreground) from the background. In contrast, semi-supervised learning methods (Bai *et al.*, 2017; Baur, Albarqouni & Navab, 2017; Min & Chen, 2018; Zhou *et al.*, 2019b; Perone, Ballester, Barros & Cohen-Adad, 2019) seek to improve the training of segmentation models by leveraging unlabeled images, in addition to labeled ones. Unlike weakly-supervised approaches, these methods rely on intrinsic properties of the data distribution (or *priors*) which are not specific to individual images. Semi-supervised methods for segmentation include techniques based on self-training (Bai *et al.*, 2017), model-based or data-based distillation (Gupta, Hoffman & Malik, 2016; Radosavovic, Dollár, Girshick, Gkioxari & He, 2018; Zhou *et al.*, 2019b), attention learning (Min & Chen, 2018), adversarial learning (Souly, Spampinato & Shah, 2017; Hung, Tsai, Liou, Lin & Yang, 2018; Zhang *et al.*, 2017c; Luc, Couprie, Chintala & Verbeek, 2016), and manifold embedding (Baur *et al.*, 2017).

Co-training is one of the most popular general-purpose techniques for semi-supervised learning. This technique originally proposed by Blum and Mitchell (Blum & Mitchell, 1998) is based on the idea that training examples can be described by two complementary (conditionally independent given the corresponding class labels) sets of features, called views. Multi-view learning (Xu, Tao & Xu, 2013) extends this idea to multiple complementary views. The general principle of this type of method is to simultaneously train classifiers for each view, using the labeled data, such that their predictions agree for unlabeled examples. Enforcing this agreement between classifiers reduces the search space and thus helps find a model which will generalize well to unseen data. While co-training and learning methods have been used with great success in natural language processing (Wan, 2009; Nigam & Ghani, 2000; Maeireizo, Litman & Hwa, 2004), their application to visual tasks has so far been limited (Levin, Viola & Freund, 2003). One of the main reasons for this is that such methods require complementary models to learn from independent features. Although such independent features may be available in specific scenarios (e.g., multiplanar images (Zhou *et al.*, 2019b)), there is no effective way to construct these sets from individual images. Recently, Qiao *et al.* proposed a deep co-training method for semi-supervised image recognition (Qiao, Shen, Zhang, Wang & Yuille, 2018). The main innovation of this work is to use adversarial examples, built from both labeled and unlabeled images, for imposing diversity among the different classifiers. Specifically, during training, a classifier is encouraged to output predictions similar to those of the other classifier for adversarial examples, hence classifiers will tend to disagree for those examples.

Until now, deep co-training has been applied only to classification. In contrast, semantic segmentation is a more complex problem with a larger and structured output space. In this work we extend and adapt the co-training approach for this task. The contributions of our work are as follows:

- We present a deep adversarial co-training method for semantic segmentation, extending the work of Qiao *et al.* to this more challenging problem. To our knowledge, this is the first co-training method proposed for single-image semantic segmentation.
- We show key differences between the application of deep co-training for classification and segmentation, and explore the effect of adversarial training on the prediction diversity of segmentation models.
- We conduct a comprehensive set of experiments which demonstrate the potential of co-training for segmenting different types of images. Our experiments also analyze the impact

of various elements of the method, including the number of classifiers, the trade-off between model agreement and diversity, and the generation of adversarial examples. We believe these experiments can be of benefit to future investigations on co-training methods for segmentation.

The rest of this paper is as follows. In the next section, we give a brief summary of related literature, focusing on recently proposed methods for semi-supervised segmentation. In Section 4.4, we present our deep adversarial co-training approach for segmentation. We then evaluate our method on the tasks of segmenting cardiac structures, spine and spleen in Section 5. Finally, we conclude with a summary of our contribution and results.

3.3 Related work

Semi-supervised learning has a long history in machine learning. The first methods were presented around 50 years ago for estimating mixture models (Cooper & Freeman, 1970; Dempster, Laird & Rubin, 1977). Since then, many different approaches have been proposed. Here, we will focus mostly on the most recent and promising methods for visual recognition and, more specifically, semantic segmentation. For a complete review of semi-supervised methods, see (Chapelle, Schlkopf & Zien, 2010).

A quite simple, yet powerful approach for semi-supervised learning is to select the most likely label of the current model as ground truth for unsupervised data. This is often referred to as pseudo-label (Lee, 2013) or entropy regularization (Grandvalet & Bengio, 2006). More sophisticated approaches make use of unlabeled samples, leveraging the unsupervised representation of an autoencoder (Rasmus, Berglund, Honkala, Valpola & Raiko, 2015) or a variational autoencoder (Kingma, Mohamed, Jimenez Rezende & Welling, 2014). Another line of research for semi-supervised learning is based on the idea that the pseudo-labeling can be improved and made more robust if multiple models are used for generating the pseudo-labels (Laine & Aila, 2017;

Tarvainen & Valpola, 2017). Regularizing the learning with adversarial examples is also a promising technique. It consists in generating samples that are adversarial to the model (Goodfellow, Shlens & Szegedy, 2015), i.e. samples that the model cannot classify correctly, and adding them to the training data to improve robustness. Recently, the generation of adversarial samples has been applied to unlabeled samples, therefore extending their use to semi-supervised learning with very promising results (Miyato, Maeda, Koyama & Ishii, 2018). This technique has also been used for co-training multiple classification models (Qiao *et al.*, 2018). Our proposed method is based on the last approach, but adapted to the more challenging task of semi-supervised image segmentation. For an updated evaluation of state-of-the art semi-supervised methods for image classification, see (Oliver, Odena, Raffel, Cubuk & Goodfellow, 2018).

Semi-supervised learning has also been used for image segmentation (Bai *et al.*, 2017; Baur *et al.*, 2017; Min & Chen, 2018; Zhou *et al.*, 2019b). As for classification, the main idea of semi-supervised segmentation methods is to propagate the labels of training samples to unlabeled images. However, in the case of segmentation, the output is structured and thus methods based on local vicinity of the sample representation would not work. A common approach is to use an iterative two steps procedure in which: *i*) the unlabeled images are annotated considering the output of the segmentation network as ground truth; *ii*) the network parameters are updated based on the segmented (annotated) images (Bai *et al.*, 2017). A common problem of such approach is that initial errors might be propagated and amplified to unlabeled images, producing catastrophic results. Various approaches are used to avoid this problem. For instance, model-based (Gupta *et al.*, 2016) and data-based (Radosavovic *et al.*, 2018; Zhou *et al.*, 2019b) distillation can reduce the error propagation by aggregating the prediction of multiple teacher models to train a student model (Min & Chen, 2018). Another approach proposed by Baur *et al.* (Baur *et al.*, 2017) embeds the network representation in a manifold, such that images having similar characteristics are near to each other.

Methods based on generative adversarial networks (GANs) (Goodfellow *et al.*, 2014) have recently shown promising results for semi-supervised segmentation (Souly *et al.*, 2017; Hung *et al.*, 2018; Zhang *et al.*, 2017c; Luc *et al.*, 2016). The first approach using GANs for semantic segmentation was proposed by Luc *et al.* (Luc *et al.*, 2016) and extended to the semi-supervised case in (Zhang *et al.*, 2017c). In this work, a discriminator network should distinguish between the segmentation of labeled and unlabeled images. This forces the segmentation model to perform as well on unlabeled images in order to fool the discriminator. An improved strategy is proposed by Hung *et al.* (Hung *et al.*, 2018), where the discriminator is used to predict areas of high confidence on unlabeled images. These areas are then used to update the segmentation network. It is important to distinguish GAN models from the use of adversarial examples (Goodfellow *et al.*, 2015). While GAN models are based on the simultaneous learning of two adversarial networks (the discriminator and the generator), adversarial training proposes the generation of samples with subtle modifications that can fool a learned model. Although GANs have already been employed for improving semi-supervised approaches, adversarial samples have not yet been applied to segmentation. In this paper, we show how to leverage adversarial samples in semi-supervised segmentation by exploiting a co-training procedure (Qiao *et al.*, 2018).

3.4 Methodology

3.4.1 Problem formulation

As a dense prediction problem with complex output space, semantic segmentation is extremely challenging in a semi-supervised setting. In real-life applications, particularly those related to medical imaging, such a setting is however common since manual annotation is often an expensive and time-consuming process. Consequently, only a small fraction of images in the dataset can have full pixel-wise labels. The proposed method aims to exploit both labeled and

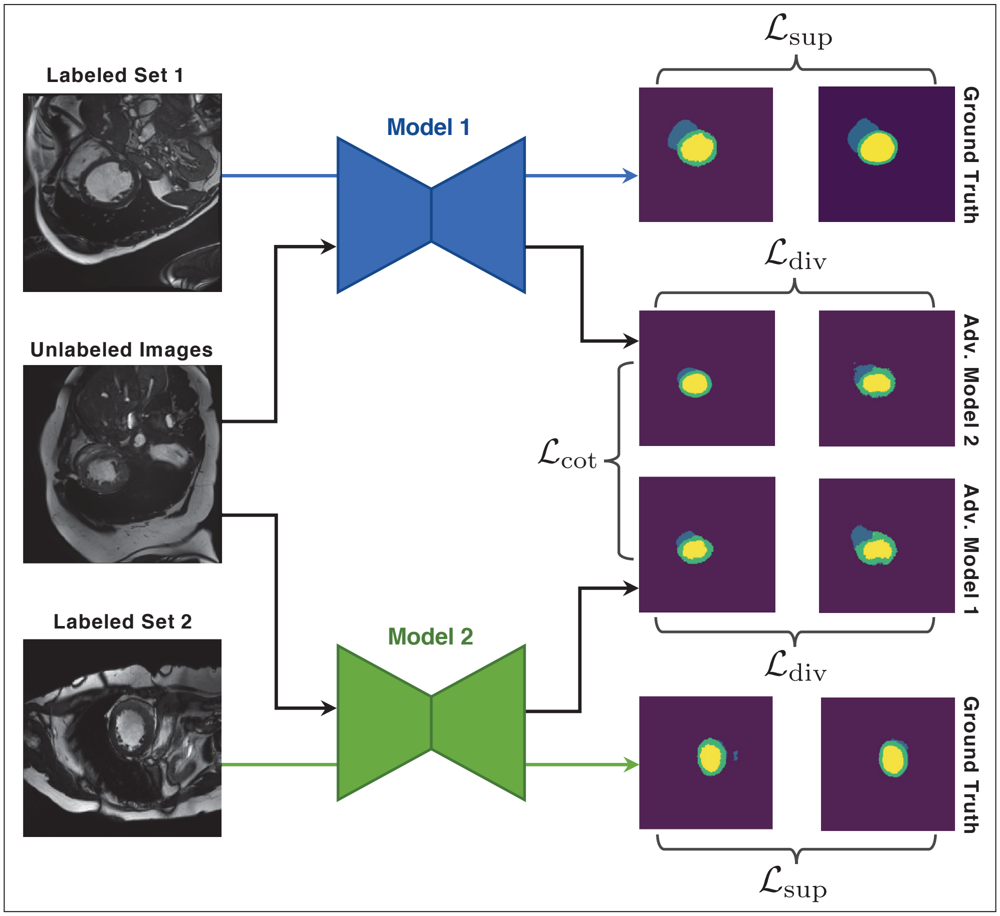

$\mathcal{L}_{\text{sup}}$
Labeled Set 1
Model 1
Ground Truth
$\mathcal{L}_{\text{div}}$
Unlabeled Images
Adv. Model 2
$\mathcal{L}_{\text{cot}}$
Adv. Model 1
Labeled Set 2
Model 2
$\mathcal{L}_{\text{div}}$
Ground Truth
$\mathcal{L}_{\text{sup}}$

Figure 3.1 Overview of the deep co-training approach proposed for image segmentation (dual-view setting). Two deep CNN models are trained simultaneously with different sets of labeled images and a common set of unlabeled images. The loss function is composed of three terms: $\mathcal{L}_{\text{sup}}$, $\mathcal{L}_{\text{cot}}$ and $\mathcal{L}_{\text{div}}$. Term $\mathcal{L}_{\text{sup}}$ ensures that network predictions for labeled examples are consistent with ground truth segmentation masks; $\mathcal{L}_{\text{cot}}$ forces networks to agree with each other for unlabeled examples; $\mathcal{L}_{\text{div}}$ imposes a network to agree with the predictions of the other network's adversarial examples.

unlabeled images by using the general, yet powerful principle of multi-view co-training.

We formalize the problem of image segmentation as follows. Given a set of labeled data $\mathcal{S} = \{(x_1, y_1), \ldots, (x_m, y_m)\}$, each example comprised of an image $x_i : x \to \mathcal{F}$ and corresponding ground truth segmentation mask $y : \Omega \to C$, where Ω is the set of image pixels (or voxels in the 3D case), $\mathcal{F}$ the set of pixel features (e.g., $\mathcal{F} = \mathbb{R}$ for grey-scale images), and C the set of possible labels. In a semi-supervised setting, we also have a set of n unlabeled images $\mathcal{U} = \{x_1, \ldots, x_n\}$, with $n \gg m$, without ground truth labels. The goal is to learn from $\mathcal{D} = \mathcal{S} \cup \mathcal{U}$ a segmentation model f parametrized by θ, which maps each pixel of an input image to its correct label.

3.4.2 Proposed approach

As in standard multi-view learning approaches, we train multiple models in a collaborative manner and, once trained, combine their outputs to predict the labels of new images. Motivated by the outstanding performance of deep convolutional network networks (CNNs) for various segmentation tasks (Litjens *et al.*, 2017; Dolz *et al.*, 2018a; Dolz *et al.*, 2018b), we employ this type of model in the proposed approach. Specifically, we train an ensemble of k segmentation networks $f^i(\cdot\,;\,\theta^i)$, $i = 1, \ldots, k$. We assume the network uses a softmax function at each image pixel to compute label probabilities, and denote as f^i_{jc} the probability of label c for pixel j, predicted by Model i. Without loss of generality, in what follows, we will consider a dual view setting (i.e., $k = 2$) and describe how this setting can be naturally extended to multiple views.

Following co-training methods for classification, we employ a loss function composed of a weighted sum of three separate terms to train the ensemble's segmentation models (see Fig. 3.1):

$$\mathcal{L}(\theta;\, \mathcal{D}) \;=\; \mathcal{L}_{\mathrm{sup}}(\theta;\, \mathcal{S}) \;+\; \lambda_{\mathrm{cot}}\, \mathcal{L}_{\mathrm{cot}}(\theta;\, \mathcal{U}) \;+\; \lambda_{\mathrm{div}}\, \mathcal{L}_{\mathrm{div}}(\theta;\, \mathcal{D}). \tag{3.1}$$

The three loss terms are explained in following subsections.

Supervised loss

The first term, $\mathcal{L}_{\mathrm{sup}}$, is the supervised loss obtained from labeled examples. It aggregates the loss computed separately for each model:

$$\mathcal{L}_{\mathrm{sup}}(\theta;\, \mathcal{S}) \;=\; \mathcal{L}^1_{\mathrm{sup}}(\theta^1;\, \mathcal{S}^1) + \mathcal{L}^2_{\mathrm{sup}}(\theta^2;\, \mathcal{S}^2). \tag{3.2}$$

Here, labeled data subsets $\mathcal{S}^i \subset \mathcal{S}, i \in \{1, 2\}$ can differ across models to ensure their diversity. While any segmentation loss can be considered, in this work, we employed the well-known pixel-wise cross-entropy loss, defined as

$$\mathcal{L}^i_{\text{sup}}(\theta^i; \mathcal{S}^i) \;=\; \mathbb{E}_{(x,y)\in\mathcal{S}^i}\left[\sum_{j\in x}\sum_{c\in C} y_{jc}\log f^i_{jc}(x;\theta^i)\right], \tag{3.3}$$

where $y_{jc} = 1$ if the true label of pixel j is c, else $y_{jc} = 0$ (i.e., one-hot label encoding). Supervised loss $\mathcal{L}_{\text{sup}}$ encourages models to output consistent predictions with respect to their ground truth labels.

Ensemble agreement loss

In addition to exploiting labeled information, unlabeled image dataset $\mathcal{U}$ is also used to guide the learning process. Based on the consensus principle (Xu *et al.*, 2013), we want the segmentation networks to output similar predictions for the same unlabeled images. We argue that enforcing this agreement helps improve the generalization of individual models by restricting their parameter search space to cross-view consistent solutions. Toward this goal, we minimize the distance between the class distributions predicted by different models. To make our approach compatible with more than two views, we define the agreement loss $\mathcal{L}_{\text{cot}}$ as the Jensen-Shannon divergence (JSD), which is the average Kullack-Liebler divergence D_{KL} between the prediction of each model f^i and their mean prediction $\overline{f}$:

$$\begin{aligned}\mathcal{L}_{\text{cot}}(\theta; \mathcal{U}) \;&=\; \mathbb{E}_{x\in\mathcal{U}}\left[D_{\text{KL}}\Big(f^1(x;\,\theta^1)\;||\;\overline{f}(x;\,\theta)\Big) \;+\; D_{\text{KL}}\Big(f^2(x;\,\theta^2)\;||\;\overline{f}(x;\,\theta)\Big)\right]\\ &=\; \mathbb{E}_{x\in\mathcal{U}}\left[\mathcal{H}\Big(\frac{1}{2}(f^1(x;\,\theta^1)+f^2(x;\,\theta^2))\Big) \;-\; \frac{1}{2}\Big(\mathcal{H}(f^1(x;\,\theta^1)) \;+\; \mathcal{H}(f^2(x;\,\theta^2))\Big)\right].\end{aligned} \tag{3.4}$$

In this equation, $\mathcal{H}(\cdot)$ corresponds to the Shannon entropy. Unlike KL divergence, the JSD between different distributions is symmetric, and thus loss $\mathcal{L}_{\text{cot}}$ considers the prediction of all models equally important when minimizing their disagreement.

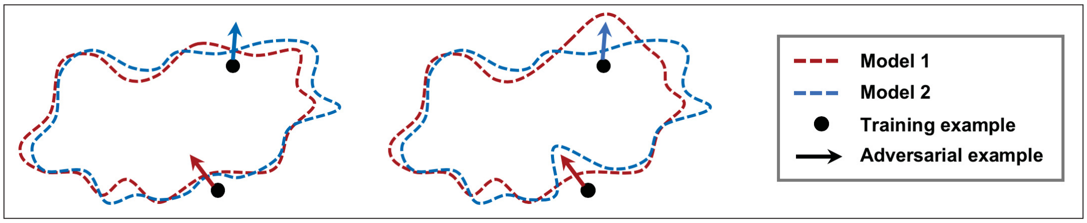


Figure 3.2 Illustration of the ensemble diversity strategy based on adversarial training. Adversarial examples are generated from training images (black dots), for both models (red and blue arrows). Each model is then forced to agree with the prediction of the other model for its own adversarial examples (right-side image).

Diversity loss

A key principle of ensemble learning is having diversity between models in the ensemble. If all models learn the same class distribution, then combining their output will not be superior to individual model predictions. In co-training, diversity is essential so that models can learn from one another during training. The standard approach for obtaining diversity is to have independent sets of features (i.e., views), or generating them by splitting available features into complementary subsets. In deep CNN classification, however, the internal representation of images is learned by the network during training, therefore such standard approach cannot be applied. Instead, we define diversity based on network output, and consider two models as different if they predict sufficiently different segmentations for some given images.

Since models in the ensemble must agree for unlabeled images, and their prediction on labeled images is constrained by ground-truth segmentation masks, training images cannot be used directly to impose diversity. Instead, we use the approach proposed by Qiao *et al.* for image classification (Qiao *et al.*, 2018), and augment the dataset with adversarial examples generated from both labeled and unlabeled data. Adversarial examples for a model are used to teach other models in the ensemble. In the case of dual-view co-training, we define our diversity loss as

$$\mathcal{L}_{\text{div}}(\theta;\mathcal{D}) \;=\; \mathbb{E}_{x\in\mathcal{D}}\Big[\mathcal{H}\Big(f^1(x;\theta^1),\, f^2(g^1(x);\theta^2)\Big) + \mathcal{H}\Big(f^2(x;\theta^2),\, f^1(g^2(x);\theta^1)\Big)\Big], \tag{3.5}$$

where $\mathcal{H}(\cdot,\cdot)$ refers to cross-entropy and $g^i(x)$ is an adversarial example targeted on model $f^i(\cdot;\theta^i)$, given input image x. As illustrated in Fig. 3.2, this loss function encourages a model to be robust to the adversarial examples generated for the other one, thereby avoiding the collapse of their decision boundary on each other (i.e., the adversarial loss reaches its maximum value when the two networks are identical).

The diversity imposed by the loss can also be motivated as follows. If example $g^1(x)$ is adversarial for Model 1, then we have that $f^1(x;\,\theta^1) \neq f^1(g^1(x);\,\theta^1)$. Moreover, minimizing the first term of Eq. (3.5) will impose that $f^1(x;\,\theta^1) = f^2(g^1(x);\,\theta^2)$. Last, combining both relations yields $f^1(g^1(x);\,\theta^1) \neq f^2(g^1(x);\,\theta^2)$. Applying the same idea for Model 2, we conclude that models will disagree on adversarial examples of each model. One should note, however, that the above relations are not guaranteed to hold in practice (e.g., predictions can be very similar but not equal). In our experiments, we show that differences mostly occur on the boundary between different regions, which is where most segmentation mistakes are made (see Fig. 3.9).

Adversarial examples are generated by adding small perturbations to input images, so as to change the network's prediction as much as possible. In this work, we generate these examples using distinct schemes depending on the source of the image x. If x is drawn from the unlabeled dataset $\mathcal{U}$, we apply the Virtual Adversarial Training (VAT) (Miyato *et al.*, 2018) method because no ground truth is available. VAT optimizes local distribution smoothness (LDS) which measures the robustness of the model against virtual adversarial direction. Following VAT, we generate an adversarial example from training image x as $x_{\text{adv}} = x + r_{\text{adv}}$, where

$$r_{\text{adv}} \;=\; \underset{r;\,\|r\|_2\leq\epsilon}{\arg\max}\; D_{\text{KL}}(f(x;\,\theta)\,||\,f(x+r;\,\theta)). \tag{3.6}$$

On the other hand, when x is drawn from the labeled set $\mathcal{S}$, we instead apply the Fast Gradient Sign Method (FGSM) since it can produce noise targeted to the ground truth, thus providing more valuable information. In this case, adversarial examples x_{adv} are generated with FGSM as

$$x_{\text{adv}} \;=\; x \;+\; \epsilon \cdot \operatorname{sign}\Big(\nabla_x \mathcal{H}(f(x;\,\theta), y)\Big), \tag{3.7}$$

where $\mathcal{H}$ is the cross-entropy loss used as in full supervision, and y is the true label of x. This approach also constrains the magnitude of adversarial perturbations using a predefined ϵ parameter.

Algorithm 3.1 Deep Co-Training Segmentation (*training*)

```
Input: Labeled images $\mathcal{S} = \{(x_1, y_1), \ldots, (x_m, y_m)\}$;
Input: Unlabeled images $\mathcal{U} = \{x_1, \ldots, x_n\}$;
Input: Number of views $k$;
Output: Network parameters $\{\theta^i\}_{i=1}^k$;
Initialize network parameters $\theta^i$, $i = 1, \ldots, k$;
for epoch = $1, \ldots, E_{\max}$ do
    for iter = $1, \ldots, T_{\max}$ do
        Randomly choose two different networks $\theta^{i_1}$ and $\theta^{i_2}$;
        Draw two batches $\mathcal{S}^{i_1}, \mathcal{S}^{i_2} \subset \mathcal{S}$ of $b$ labeled images $(x, y)$ (with replacement);
        Draw a single batch $\mathcal{U}^b \subset \mathcal{U}$ of $b$ unlabeled images $x$;
        Compute adversarial examples $g^{i_1}(x)$ for all $x \in \mathcal{S}^{i_1} \cup \mathcal{U}^b$, and $g^{i_2}(x)$ for all
          $x \in \mathcal{S}^{i_2} \cup \mathcal{U}^b$, using Eq. (6) or (7);
        Let $\mathcal{L} = \mathcal{L}_{\text{sup}} + \lambda_{\text{cot}}\,\mathcal{L}_{\text{cot}} + \lambda_{\text{div}}\,\mathcal{L}_{\text{div}}$, as defined in Eq. (3.2)-(3.5), using $\mathcal{S}^{i_j}$
          for the supervised loss of model $i_j$;
        Compute gradients w.r.t. $\mathcal{L}$ and update parameters $\theta^{i_j}$, $j = 1, 2$, using
          back-propagation;
    end for
    Update learning rate and parameters $\lambda_{\text{cot}}, \lambda_{\text{div}}$ as in Eq. (3.8);
end for
return $\{\theta^i\}_{i=1}^k$ ;
```

Training and testing process

The whole training process is summarized in Algorithm 3.1. The algorithm takes as input labeled images $\mathcal{S}$, unlabeled images $\mathcal{U}$, and the number k of segmentation models to train (views). It outputs the parameters of the k trained models, i.e. $\{\theta^i\}_{i=1}^k$. At every training epoch, the algorithm performs $T_{\max}$ mini-batch iterations to update the network parameters. In each iteration, we randomly select a pair of networks to generate adversarial examples and compute the supervised, co-training and diversity loss functions. In practice, network pairs are sampled such that all networks are updated at each $\lceil k/2 \rceil$ iterations. At the end of each epoch, we modify the learning rate using standard decay, and update the co-training and diversity loss parameters λ_{cot} and λ_{div} with a dynamic strategy. This strategy follows a Gaussian ramp-up curve defined by parameters $\lambda_{\max}$, t_{ini} and t_{end}:

$$\lambda(t) = \begin{cases} 0 & \text{, if } \; t < t_{\text{ini}} \\ \lambda_{\max} \cdot \exp\left(-5 \cdot \left(1 - \frac{t - t_{\text{ini}}}{t_{\text{end}} - t_{\text{ini}}}\right)^2 \right) & \text{, if } \; t_{\text{ini}} \leq t < t_{\text{end}} \\ \lambda_{\max} & \text{, if } \; t \geq t_{\text{end}} \end{cases} , \tag{3.8}$$

An example of the ramp-up function is shown in Fig. 3.3. The ramp-up only starts after t_{ini} epochs to avoid hampering training in its early stage, and reaches and its maximum value $\lambda_{\max}$ after t_{end} epochs.

In testing, we feed an unlabeled image to the trained models and combine their outputs to obtain the final segmentation. This can be done in different ways, for instance, using hard- or soft-voting. In hard-voting, the label of a pixel is the one predicted by the majority of models (with random tie-breaking). On the other hand, soft-voting consists in averaging the pixel-wise class probabilities across models, and using this average as ensemble prediction. The latter

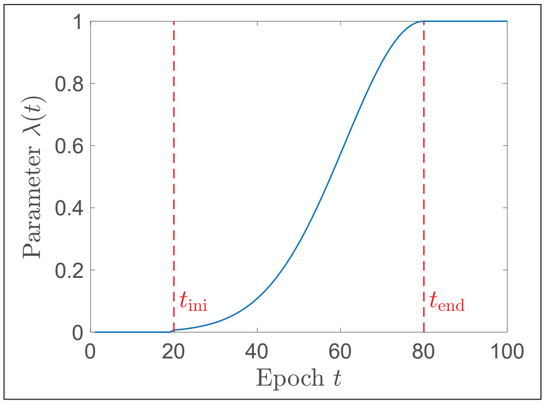


Figure 3.3 Example of ramp-up function $\lambda(t)$ for $\lambda_{\max} = 1$, $t_{\text{ini}} = 20$ and $t_{\text{ini}} = 80$.

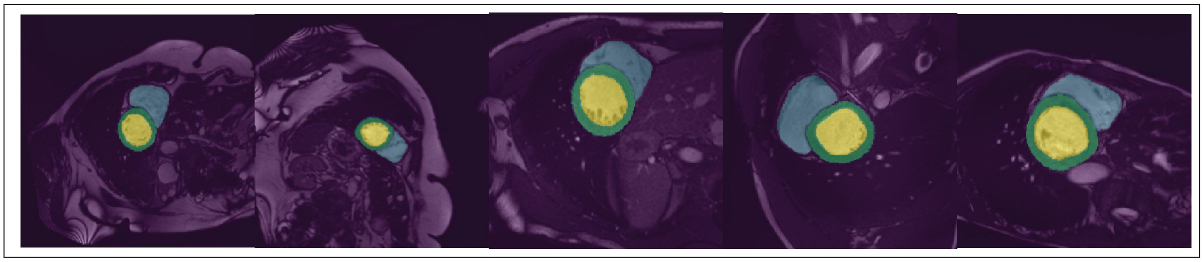


Figure 3.4 Examples of images and ground truth segmentation masks in the ACDC dataset. Images are segmented in four separate classes: endocardium of the left ventricle (LV, yellow), myocardium of the left ventricle (Myo, green), endocardium of the right ventricle (RV, blue), background (purple).

technique is commonly used in homogeneous ensemble techniques like bootstrap aggregating (bagging).

3.5 Experiments and results

3.5.1 Evaluation datasets and metrics

Our experiments are conducted on three clinically-relevant benchmark datasets for medical image segmentation: Automated Cardiac Diagnosis Challenge (ACDC) (Bernard *et al.*, 2018), Spinal Cord Gray Matter Challenge (SCGM) (Prados *et al.*, 2017), and Spleen sub-task dataset of the Medical Segmentation Decathlon Challenge (Simpson *et al.*, 2019).

- **ACDC dataset**: The publicly available ACDC dataset consists of 200 short-axis cine-MRI scans from 100 patients, evenly distributed in 5 subgroups: normal, myocardial infarction, dilated cardiomyopathy, hypertrophic cardiomyopathy, and abnormal right ventricles. 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. Segmentation masks delineate 4 regions of interest: left ventricle endocardium (LV), left ventricle myocardium (Myo), right ventricle endocardium (RV), and background (see Fig. 3.4). For our experiments, we used a split of 75 subjects (150 scans) for training and 25 subjects (50 scans) for testing. Short-axis slices within 3D-MRI scans were considered as 2D images, which were re-sized to 256×256.
- **SCGM dataset**: The SPGM dataset is a publicly-available collection of multi-center, multi-vendor MRI. It comprises a total of 80 healthy subjects (age range of 28.3 to 44.3 years) obtained by four different centers, with 20 subjects from each center. Scans were acquired using different MRI systems and distinct acquisition parameters, leading to high-variability of image characteristics: resolution range of $0.25 \times 0.25 \times 2.5$ mm to $0.5 \times 0.5 \times 5.0$ mm, number of axial slices range of 3 to 28. The training set contains 40 labeled scans, each annotated slice-wise by 4 independent experts and the ground truth mask obtained by majority voting. Ground truth labels for the remaining 40 test images are not available. For additional details on the dataset, see Prados *et al.* (2017).

  In Perone *et al.* (2019), this dataset is used to train and test a semi-supervised segmentation method based on the mean teacher algorithm. Experiments of this work, which focused on domain adaptation, used images from centers 1 and 2 as the training set, images from center 3 as the validation set, and images from center 4 as the test set. In our work, we seek to evaluate methods in a more traditional semi-supervised setting, where very few labeled images are seen in training. Hence, we consider a different training set where labeled images only come from center 1 (total of 30 images), and unlabeled images from all centers are used (total of 465 images). The test set contains labeled images from centers 3 and 4 (total of 264 images). Following (Perone *et al.*, 2019), slices in each scan are first resampled to a uniform resolution of 0.25 × 0.25 mm, and then center-cropped to a size of 200 × 200 pixels.

- **Spleen datset**: As one of the ten sub-tasks of the Medical Segmentation Decathlon Challenge (Simpson *et al.*, 2019), the publicly-available Spleen dataset[2] consists of patients undergoing chemotherapy treatment for liver metastases. A total of 61 portal venous phase CT scans (only 41 were given with ground truth) were included in the dataset with acquisition and reconstruction parameters described in Simpson *et al.* (2019). The ground truth segmentation was generated by a semi-automatic segmentation software and then refined by an expert abdominal radiologist.

  For our experiments, 2D images are obtained by slicing the high-resolution CT volumes along the axial plane, followed by a max-min normalization with a range between 0 and 1. Each slice is then resized to a resolution of 256×256 or 512×512 to test the robustness of the different algorithms to various input image resolutions. In order to evaluate these algorithms in a semi-supervised setting, we split the dataset into labeled, unlabeled and validation image subsets, comprising CT scans of 4, 32, and 5 patients respectively.

As in similar studies, we use the Dice similarity coefficient (DSC) and the Hausdorff distance (HD) to evaluate the performance of segmentation models. DSC measures the overlap between the predicted segmentation S and ground truth segmentation G:

$$\mathrm{DSC}(S, G) = \frac{2|S \cap G|}{|S| + |G|}. \tag{3.9}$$

On the other hand, HD is a boundary distance metric which measures the largest distance (in

[2] http://medicaldecathlon.com/

mm) between a point in S and its nearest point in G (or vice-versa):

$$\mathrm{HD}(S,G) \;=\; \max\,\{d(S,G),\; d(G,S)\}. \tag{3.10}$$

Unlike for DSC, where a perfect segmentation has a value of 1 and the worse possible segmentation a value of 0, a smaller HD value indicates a better segmentation.

3.5.2 Experimental details

As segmentation network, we employed the well-known U-Net (Ronneberger *et al.*, 2015) architecture, with 15 layers, Dropout and ReLU activations. This architecture is one of the most popular models for segmentation, especially for tasks related to medical imaging. The same data augmentation strategy was considered for all datasets, which applies random rotation, flip, and random crop of 85-95% surface on the original image.

Networks were trained using stochastic gradient descent (SGD) with the Adam optimizer. Learning parameters were set separately for different datasets. For the ACDC and Spleen datasets, we used a maximum number of epochs of 300, an initial learning rate of 0.001 and a weight decay of 0.0001. The learning rate was decreased by a factor of 10 every 90 epochs. Batch size was set to 4 for both labeled and unlabeled data. FSGM with $\epsilon = 0.03$ or VAT with $\epsilon = 10$ was used to create adversarial examples. For SCGM, the maximum number of epochs was set to 300, and learning rate decreased by a factor of 10 each 100 epochs. All other parameters remained the same for this dataset. For all running experiments, we used the ramp-up strategy of Eq. (3.8) to set hyper-parameters λ_{cot} and λ_{div}. We set t_{ini} to 1 for λ_{cot} and 20 for λ_{div}, since adversarial noise is meaningless if networks are not training enough. Moreover, we used $t_{\mathrm{end}} = 50$ for both λ_{cot} and λ_{div}. Last, we set the maximum hyper-parameter value $\lambda_{\max}$ to 0.5 for λ_{cot} and 0.05 for λ_{div}. Note that all hyper-parameters of our method, as well as comparison baselines described below, were selected using grid search on the validation set.

We report the average performance of individual models, as well as the performance of combining the prediction of all models using a voting strategy. In preliminary experiments, we observed that soft-voting usually outperformed hard-voting and thus only considered this strategy. Our deep co-training method is compared against three popular approaches for semi-supervised learning: the Pseudo Label algorithm (Lee, 2013), VAT (Miyato *et al.*, 2018) and Mean Teacher (Perone *et al.*, 2019). To our knowledge, Mean Teacher is the only other approach using multiple deep CNNs for semi-supervised segmentation. For these three baselines, we follow the same optimization, learning rate decay, weight scheduler, and data augmentation setting as for our method. For the Pseudo Label algorithm, we consider the α% most confident pixels of a prediction as ground truth, and increase α from 50% to 99% over training epochs. For VAT, we apply the same adversarial attack setting as in our method. For Mean Teacher, as in Perone *et al.* (2019), data augmentation is applied to input images of a student model and non-augmented images are fed to a teacher model, whose parameters θ' are computed by running an exponential moving average on the student's parameters θ:

$$\theta'_t \;=\; \alpha\theta'_{t-1} + (1-\alpha)\theta_t. \tag{3.11}$$

In our experiments, we set α to 0.99. Finally, the student's output for augmented images is forced to be consistent with the teacher's prediction, augmented using the same strategy, via an L_2 loss. We then report the performance of the teacher network.

3.5.3 Experimental results

ACDC dataset

We first evaluate our deep co-training method on the ACDC dataset using a dual view setting, i.e., training two segmentation models using the proposed loss. Performance is measured for individual models (we report their mean accuracy), as well as for the combined prediction using soft-voting. To simulate different levels of supervision, we vary the ratio l_a of labeled images in the training set, $0 \leq l_a \leq 1$. Images and ground-truth segmentation masks from the first $75 \times l_a$ training subjects are used as labeled data, while the images of remaining subjects serve as unlabeled data.

As additional baseline for an ablation study, we trained the two models independently, without considering the ensemble agreement (i.e., $\mathcal{L}_{\text{cot}}$) and adversarial diversity (i.e., $\mathcal{L}_{\text{div}}$) loss terms. In the presentation of results, this baseline is referred to as Independent. Note that the soft-voting score of this baseline corresponds to the well-known bagging technique in ensemble learning. As fully-supervised baseline, we also report the performance obtained by training a single model with all available training examples. This baseline is denoted as full supervision (Full) in results. Moreover, to measure the relative contribution of the adversarial loss terms on performance, we also give the average and soft-voting score of the ensemble trained without this term, and denote this approach as JSD in the results. Our proposed method, which combines all three loss terms, is referred to as Deep Co-Training Segmentation (DCT-Seg).

Table 3.1 gives the class-wise mean DSC and HD of tested methods for a labeled data ratio of $l_a = 0.2$. To evaluate robustness against parameter initialization, we ran the experiment three times with different random seeds, and computed the average and standard deviation of performance values over the three runs. We report both the ensemble average score (*avg* in the table) and the score obtained by ensemble soft-voting (*voting* in the table).

For both DSC and HD, ensemble soft-voting leads to a higher accuracy than the prediction of individual models, in all cases. This confirms the benefit of aggregating predictions from different models. It can also be observed that considering ensemble agreement without diversity (JSD) leads to a higher accuracy than the supervised loss alone (Independent). For DSC, combining all three losses in DCT-Seg gives the best performance, with overall mean improvements of 5.63% compared to Independent and 3.16% over Mean Teacher. With only 20% of training images

Table 3.1 DSC and HD performance of tested methods for validation images of the ACDC dataset. Except for full supervision (Full), all methods were trained with 20% of labeled data. Independent ($\mathcal{L}_{\text{sup}}$ only), JSD ($\mathcal{L}_{\text{sup}}$+$\mathcal{L}_{\text{cot}}$) and DCT-Seg ($\mathcal{L}_{\text{sup}}$+$\mathcal{L}_{\text{cot}}$+$\mathcal{L}_{\text{div}}$) were trained in a dual-view setting. For these methods, we report the average ensemble performance (avg) and the performance obtained by combining ensemble predictions with soft-voting (voting). *Note*: reported values are the average (standard deviation in parenthesis) obtained over three separate runs, each one with a different random seed.

Method		**DSC (%)**			
		RV	**Myo**	**LV**	**Mean**
Full		81.96 (0.15)	85.39 (0.20)	91.82 (0.15)	86.39 (0.10)
Pseudo Label (Lee, 2013)		74.60 (0.32)	78.91 (0.21)	85.79 (0.17)	79.77 (0.14)
VAT (Miyato *et al.*, 2018)		72.78 (0.39)	80.81 (0.21)	87.60 (0.18)	80.39 (0.15)
Mean Teacher (Perone *et al.*, 2019)		74.62 (1.10)	80.66 (0.61)	86.75 (0.27)	80.68 (0.41)
Independent	avg	68.82 (1.90)	78.30 (1.55)	85.92 (0.62)	77.68 (1.48)
	voting	68.28 (1.61)	79.94 (1.00)	86.41 (0.29)	78.21 (0.89)
JSD	avg	74.75 (1.69)	81.85 (0.42)	89.73 (0.58)	82.11 (0.44)
	voting	75.06 (1.87)	82.64 (0.57)	**90.31 (0.47)**	82.67 (0.67)
DCT-Seg (ours)	avg	77.51 (0.69)	82.43 (0.27)	89.85 (0.26)	83.26 (0.16)
	voting	**78.20 (0.70)**	**83.11 (0.20)**	90.22 (0.24)	**83.84 (0.10)**

Method		**HD (mm)**			
		RV	**Myo**	**LV**	**Mean**
Full		11.42 (1.15)	5.80 (0.98)	4.58 (0.65)	7.27 (0.50)
Pseudo Label (Lee, 2013)		18.82 (4.58)	11.95 (2.81)	10.71 (1.27)	13.83 (1.06)
VAT (Miyato *et al.*, 2018)		17.43 (3.37)	8.60 (1.20)	8.79 (0.52)	11.61 (0.40)
Mean Teacher (Perone *et al.*, 2019)		16.12 (1.12)	7.86 (0.78)	7.41 (0.57)	10.46 (0.36)
Independent	avg	21.26 (3.04)	13.31 (2.17)	9.21 (1.95)	14.59 (1.87)
	voting	8.77 (2.18)	6.65 (1.83)	5.00 (1.13)	6.81 (0.72)
JSD	avg	15.84 (1.59)	7.18 (0.47)	5.63 (0.88)	9.55 (0.30)
	voting	**7.39 (0.77)**	4.21 (0.33)	**3.33 (0.12)**	**4.97 (0.04)**
DCT-Seg (ours)	avg	16.05 (2.19)	7.89 (1.67)	4.98 (0.59)	9.64 (0.30)
	voting	7.43 (0.62)	**4.19 (0.29)**	**3.33 (0.09)**	4.98 (0.10)

labeled, DCT-Seg provides a mean DSC only 2.55% less than full supervision. With respect to HD, our DCT-Seg method outperformed all three baselines by a significant margin. However, enforcing model diversity did not lead to noticeable improvements in this case, with DCT-Seg achieving a performance similar to JSD. This can potentially be explained by the fact that HD is

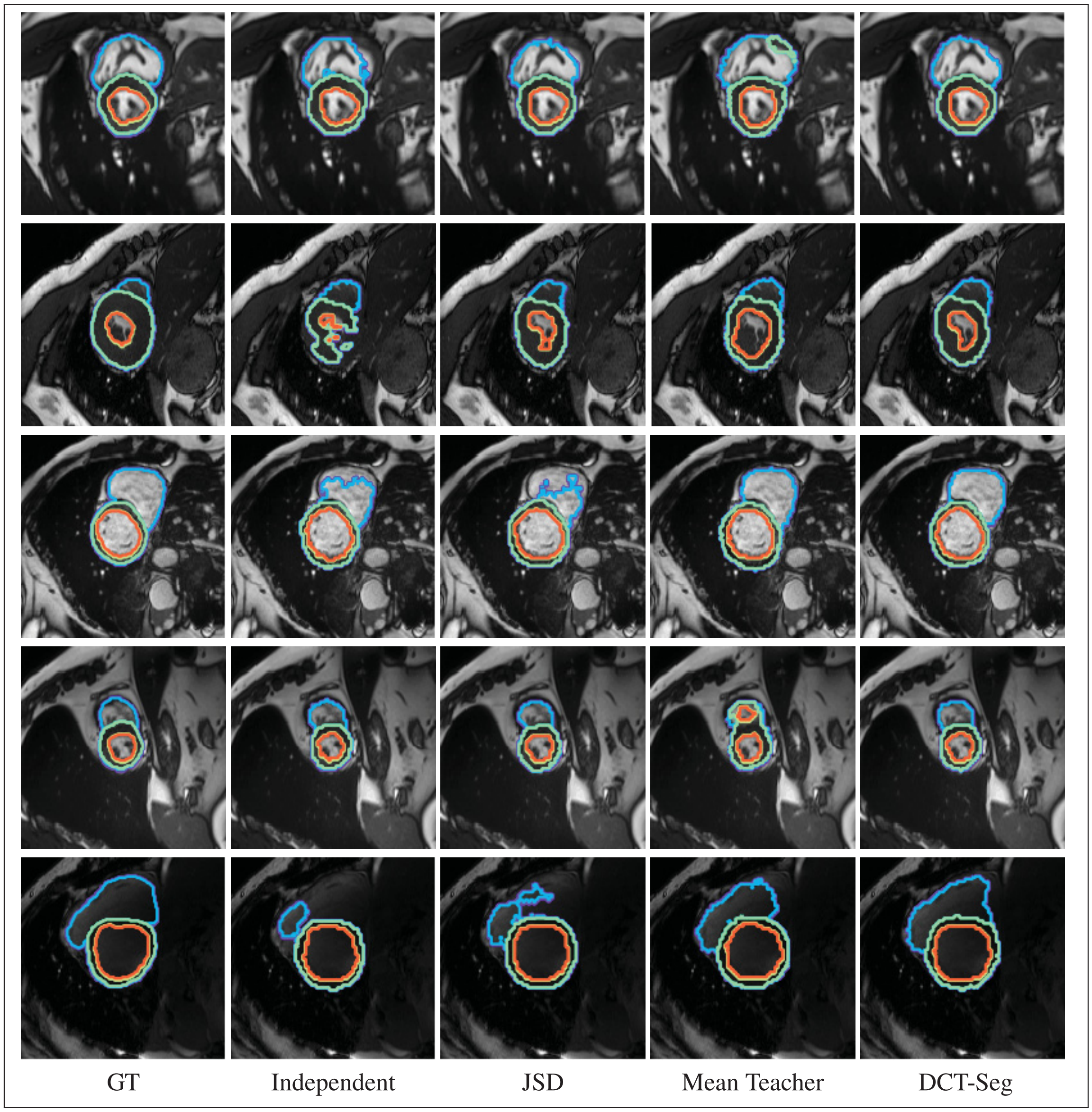


Figure 3.5 Examples of segmentation results for the ACDC dataset with 20% of labeled training examples. From left to right: Ground-truth (GT), Independent ($\mathcal{L}_{\text{sup}}$ only), JSD ($\mathcal{L}_{\text{sup}}$+$\mathcal{L}_{\text{cot}}$), Mean Teacher (Perone *et al.*, 2019), and our DCT-Seg method ($\mathcal{L}_{\text{sup}}$+$\mathcal{L}_{\text{cot}}$+$\mathcal{L}_{\text{div}}$).

more sensitive to outliers that can result from adversarial training. Examples of segmentation results for tested methods are shown in Fig. 3.5. We see that deep co-training gives contours closer to the ground-truth, with very few artifacts on the boundaries between different regions.

Table 3.2 DSC performance on the ACDC validation set when training different numbers of segmentation models (i.e., views) separately (Independent) or with the proposed deep co-training method (DCT-Seg). In this experiment, 20% of training images are labeled. *Note*: reported values are the average (standard deviation in parenthesis) obtained over three separate runs, each one with a different random seed.

Method		**2 views**	**3 views**	**4 views**
Independent	avg	77.68 (1.48)	77.80 (1.27)	77.82 (0.92)
	voting	78.21 (0.89)	78.57 (0.71)	79.08 (1.21)
DCT-Seg (ours)	avg	83.26 (0.16)	83.80 (0.38)	83.43 (0.24)
	voting	**83.84 (0.10)**	**84.71 (0.53)**	**84.61 (0.28)**

Next, we assess whether having more models in the ensemble (i.e., more than two views) can further boost performance of methods. Toward this goal, we repeated the experiment with 2, 3 and 4 views, once more using a labeled image ratio of $l_a = 0.2$. The overall mean DSC of tested methods, computed over the all classes, is reported in Table 3.2. We see that increasing the number of views does not significantly improve the performance for individually-trained models (Independent). On the other hand, for deep co-training, a small increase in DSC is observed when going from 2 to 3 views. However, adding a fourth view does not further improve performance, suggesting that co-training can effectively capture variability with a very limited number of views.

Table 3.3 DSC performance on the ACDC validation set when training two segmentation models separately (Independent) or with the proposed deep co-training method (DCT-Seg), for three different ratios l_a of labeled examples. *Note*: reported values are the average (standard deviation in parenthesis) obtained over three separate runs, each one with a different random seed.

Method		$l_a = 5\%$	$l_a = 10\%$	$l_a = 20\%$	$l_a = 50\%$
Independent	avg	69.72 (0.10)	74.68 (0.58)	77.68 (1.48)	84.96 (0.13)
	voting	71.17 (0.19)	75.84 (0.49)	78.21 (0.89)	85.12 (0.08)
DCT-Seg (ours)	avg	77.81 (0.10)	82.36 (0.33)	83.26 (0.16)	86.02 (0.14)
	voting	**78.17 (0.12)**	**82.90 (0.22)**	**83.84 (0.10)**	**86.15 (0.09)**

As third experiment, we evaluate how the proportion of labeled data impacts results in a dual-view setting. Table 3.3 gives the performance of individually-trained models (Independent) and co-training for three labeled data ratio: 10%, 20% and 50%. A clear trend is observed in these results, where mean DSC values increase sharply with the ratio of labeled images in training. In all cases, deep co-training leads to a higher DSC than training models separately, the most significant improvements obtained for the smallest ratios of $l_a = 0.05$ (7.00%) and $l_a = 0.1$ (7.06%).

SCGM dataset

To further validate the effectiveness of our proposed deep co-training method, we evaluated it on the task of segmenting spinal cord grey matter in images from the SCGM dataset. As mentioned previously, this experiment aims at testing our method in a challenging setting where very few labeled images are used in training (i.e., only 30 images), and test images are generated using different acquisition parameters.

Table 3.4 DSC performance of tested methods for validation images of the SCGM dataset. Independent ($\mathcal{L}_{\text{sup}}$ only), JSD ($\mathcal{L}_{\text{sup}}$+$\mathcal{L}_{\text{cot}}$) and DCT-Seg ($\mathcal{L}_{\text{sup}}$+$\mathcal{L}_{\text{cot}}$+$\mathcal{L}_{\text{div}}$) were trained in a dual-view setting. For these methods, we report the average ensemble performance and the DSC obtained by combining ensemble predictions with soft-voting. *Note*: reported values are the average from two separate runs, each one with a different random seed.

Method		**DSC**
Pseudo Label (Lee, 2013)		60.03
VAT (Miyato *et al.*, 2018)		59.40
Mean Teacher (Perone *et al.*, 2019)		50.55
Independent	avg	43.31
	voting	43.22
JSD	avg	45.59
	voting	44.96
DCT-Seg (ours)	avg	71.09
	voting	**72.76**

Results of this experiment are summarized in Table 3.4. Important differences can be observed between the DSC of tested methods. In this case, JSD improves the results of Independent only slightly, while our deep co-training approach outperforms both these methods by nearly 25%. This suggests that adversarial learning is highly useful when supervised training is limited (i.e., few labeled training examples, different from test examples). Compared to other tested semi-supervised approaches, our DCT-Seg method gives a mean DSC 12% higher than the best baseline (Pseudo Label). The better accuracy of deep co-training can be appreciated in Fig. 3.6, which shows examples of segmentation results for tested methods.

Table 3.5 DSC performance of tested methods for validation images of the Spleen dataset with resolutions of 256×256 and 512×512. Independent ($\mathcal{L}_{\text{sup}}$ only), JSD ($\mathcal{L}_{\text{sup}}$+$\mathcal{L}_{\text{cot}}$) and DCT-Seg ($\mathcal{L}_{\text{sup}}$+$\mathcal{L}_{\text{cot}}$+$\mathcal{L}_{\text{div}}$) were trained in a dual-view setting. For these methods, we report the average ensemble performance and the DSC obtained by combining ensemble predictions with soft-voting. *Note*: reported values are the average from two separate runs, each one with a different random seed.

Method		**DSC (%)**	
		256×256	512×512
Pseudo Label (Lee, 2013)		85.71	84.83
VAT (Miyato *et al.*, 2018)		86.82	87.16
Mean Teacher (Perone *et al.*, 2019)		86.87	87.55
Independent	avg	84.71	86.63
	voting	86.21	89.35
JSD	avg	87.92	90.04
	voting	88.96	90.73
DCT-Seg (ours)	avg	89.30	91.06
	voting	**90.19**	**91.81**

Spleen dataset

We then investigate the robustness of our proposed algorithm to different data modalities and input resolutions. Toward this goal, we repeated our experiments on the Spleen dataset consisting of 2D slices of CT scans resized to a resolution of 256×256 or 512×512. Table 3.5 summarizes the experimental results. We see that, regardless the input image size, our proposed method

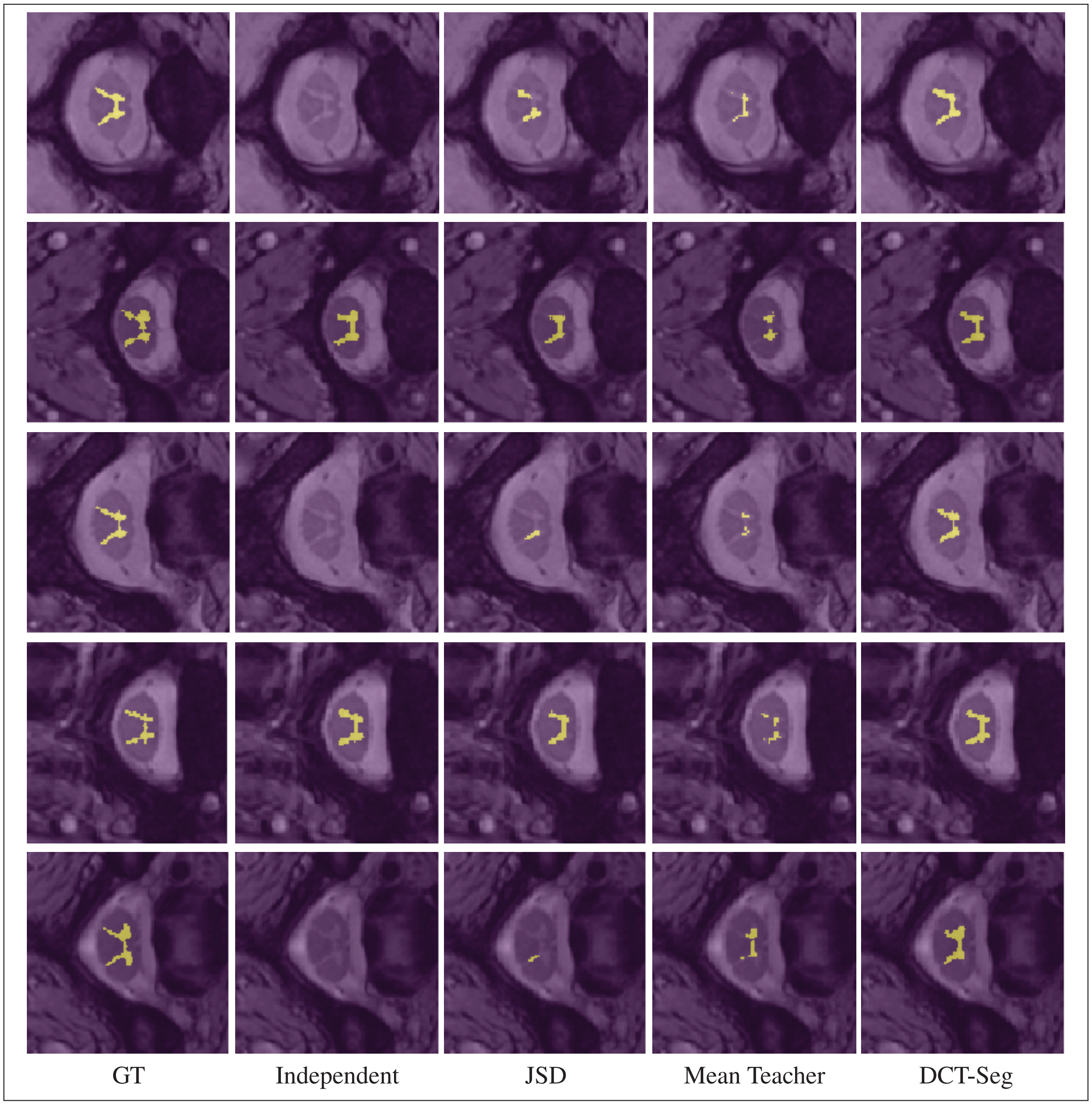


Figure 3.6 Examples of segmentation results for the SCGM dataset using Center 1 as training data. From left to right: Ground-truth (GT), Independent ($\mathcal{L}_{\text{sup}}$ only), JSD ($\mathcal{L}_{\text{sup}}$+$\mathcal{L}_{\text{cot}}$), Mean Teacher (Perone *et al.*, 2019), and our DCT-Seg method ($\mathcal{L}_{\text{sup}}$+$\mathcal{L}_{\text{cot}}$+$\mathcal{L}_{\text{div}}$).

achieves a consistent improvement over other semi-supervised approaches. Specifically, the soft-voting version of DCT-Seg obtains a mean DSC boost of 3-4% compared to the best performing baseline (Mean Teacher), showing its advantage for different image modalities and

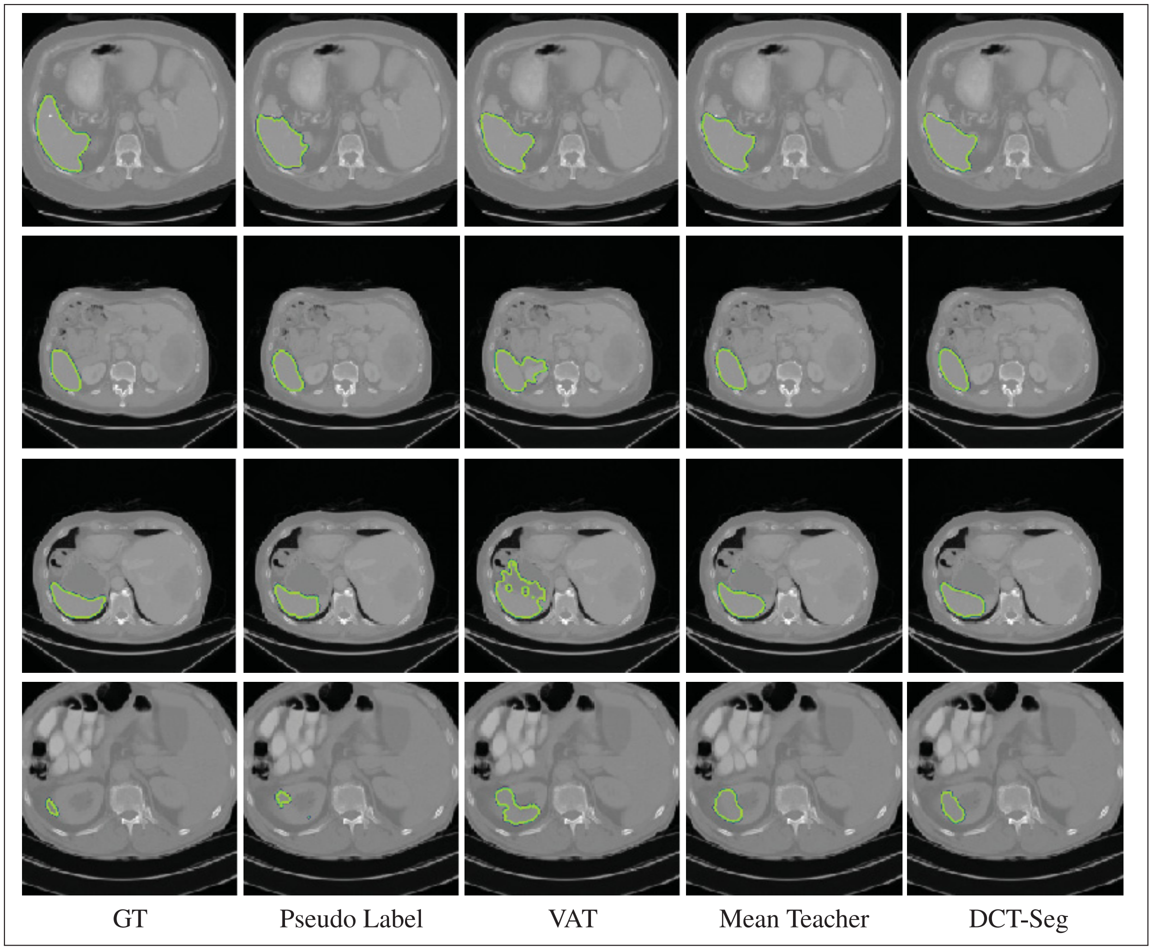
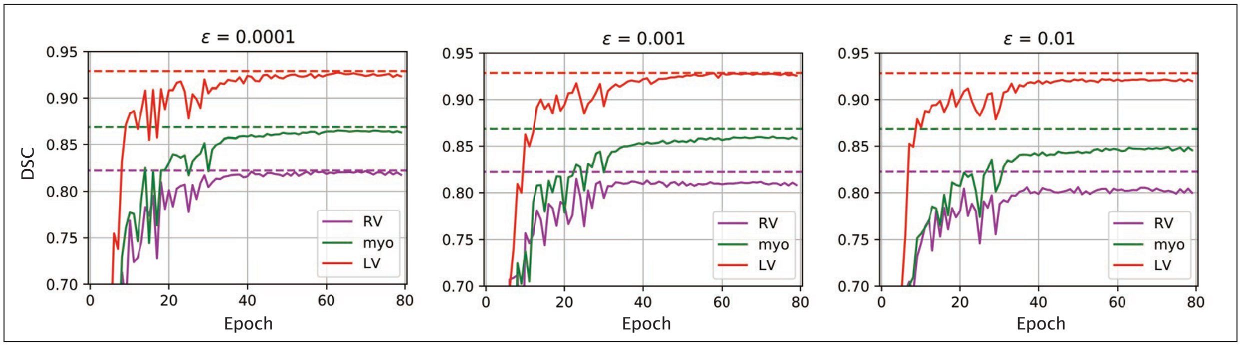


Figure 3.7 Examples of segmentation results of tested methods on the Spleen dataset with resolution of 256×256. Note: our DCT-Seg method combines the predictions of two CNNs trained with the same subset of labeled examples as other approaches.

resolutions. Examples of segmentation results obtained by tested methods on images of size 256×256 are given in Fig. 3.7. Visually, DCT-Seg and Mean Teacher provide similar results, with most pronounced differences observed for small foreground regions (e.g., last row of the figure).

3.5.4 Impact of diversity loss

We investigate the role of the ensemble diversity loss (i.e., $\mathcal{L}_{\text{div}}$) in our deep co-training method and experimentally show that it also acts as a coarse measure of model agreement, merging the prediction of models while avoiding them to collapse on each other. We perform our investigation on the ACDC dataset using two models. The first one is pre-trained by full supervision as a fixed reference, and the second one trained from scratch using a labeled data ratio of $l_a = 0.5$. Note that the trained model is only linked to the fixed reference by $\mathcal{L}_{\text{div}}$, and no supervised loss is considered while training this model. Moreover, to measure the impact of adversarial noise ϵ in $\mathcal{L}_{\text{div}}$, we repeat training with different values for ϵ.


ε = 0.0001

ε = 0.001

ε = 0.01

Figure 3.8 DSC score for models trained from scratch using only $\mathcal{L}_{\text{div}}$ with different ϵ. It can be seen that $\mathcal{L}_{\text{div}}$ acts as a similarity loss, especially when ϵ is small.

Fig. 3.8 gives the DSC obtained on the validation set by the reference model (dashed line) and model trained from scratch (solid line), for increasing amounts of adversarial noise ϵ. It can be observed that the trained model rapidly converges to the reference, without the need for a supervised signal or specific agreement loss. However, upon convergence, we see that the trained model does not fully reach the accuracy of the reference model, and that the gap between the two models is proportional to the value of ϵ. For example, a gap of 1.43%, 1.38% and 0.02% is obtained for the Myo class, when using an ϵ of 0.01, 0.001, and 0.0001, respectively. This can be explained by the fact that, when ϵ is small, adversarial examples are very similar to original images, and $\mathcal{L}_{\text{div}}$ then acts as a symmetric KL loss between the two models.

We then tested the behavior of $\mathcal{L}_{\text{div}}$ when models are trained simultaneously. Toward this goal, we initialized the two models using the same fully-supervised checkpoint and linked them only using $\mathcal{L}_{\text{div}}$. Thus, the models give the same predictions at the beginning of training. As training progresses, $\mathcal{L}_{\text{div}}$ is minimized and the models should become different from one another. We show this tendency by imposing a small $\epsilon = 0.001$ during training. With the decrease of $\mathcal{L}_{\text{div}}$, differences start appearing along region boundaries, leading to slightly worse DSC scores. Examples of prediction disagreement, measured by the L_1 norm, are shown in Fig. 3.9. It can be observed that most prediction differences occur at the boundary and within regions which are hardest to segment (i.e., left ventricle myocardium and right ventricle endocardium).

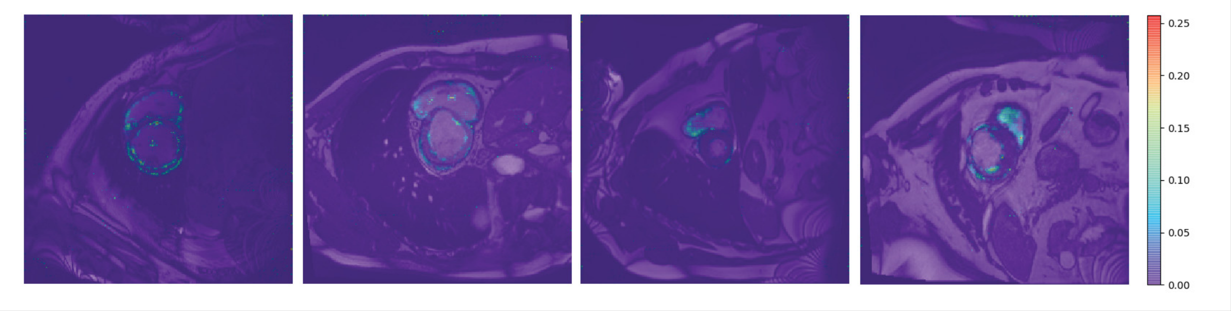


Figure 3.9 Examples of prediction disagreement between two models linked with the ensemble diversity loss ($\mathcal{L}_{\text{div}}$), measured using L_1 norm.

Last, we illustrate in Fig. 3.10 the effect of adversarial examples on model prediction diversity. For an input image, the two models can offer similar predictions. However, if this image is modified using adversarial noise, the predictions of the two models can differ significantly from one another. This confirms the usefulness of adversarial training for generating diversity between models.

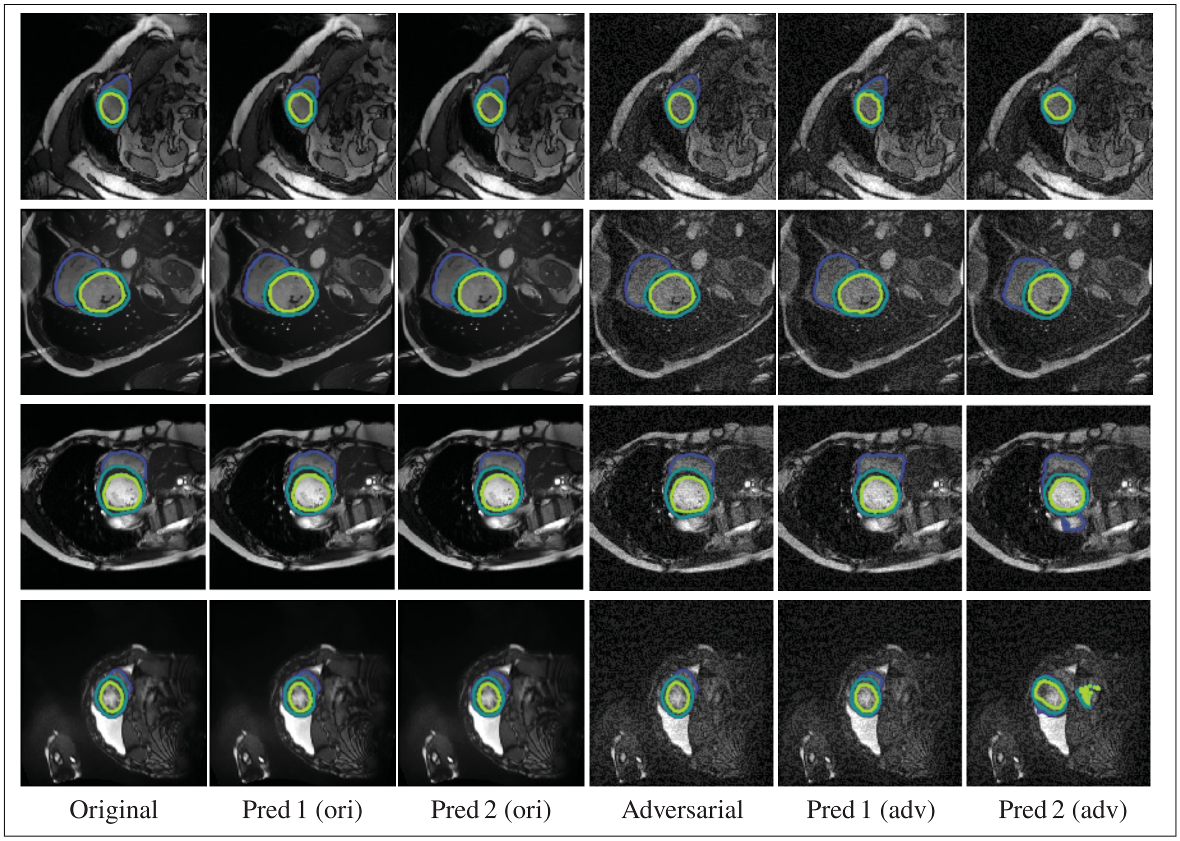


Figure 3.10 Impact of adversarial noise on prediction diversity. From left to right: original image (with GT contours), predictions of Models 1 and 2 for the original image, adversarial image for Model 2 (with GT contour), and predictions of Model 1 and 2 for the adversarial image.

3.6 Discussion and conclusion

We proposed the first application of deep co-training to single image segmentation and demonstrated its usefulness on three public benchmark datasets. Our experiments showed that both ensemble agreement and diversity loss terms help boost performance compared to standard techniques such as bagging, and that combining both in a deep co-training algorithm outperforms recent approaches like Pseudo Label, VAT and Mean Teacher.

A limitation of the proposed method is the need to train multiple segmentation networks at the same time, which increases the computational requirements and restricts the number of views possible. During testing, computing and combining multiple segmentation predictions also entails greater computational resources, although these predictions can be obtained in parallel (e.g., on separate GPUs). Nevertheless, our experiments on the ACDC dataset suggest that increasing the number of segmentation models beyond two offers limited benefits, showing the ability of our diversity-inducing strategy to capture variability in the data.

Another possible drawback of our method is the need to balance three different loss terms (i.e., $\mathcal{L}_{\text{sup}}$, $\mathcal{L}_{\text{cot}}$ and $\mathcal{L}_{\text{div}}$) that can compete against one another during training. To alleviate this problem, we proposed a ramp-up strategy where a greater importance is given to the supervised loss in initial training epochs. However, this strategy still requires some tuning which can affect performance. A useful extension of this work could be to investigate self-tuning mechanisms which can adapt more efficiently to new datasets.

In this work, an adversarial learning technique was employed to enforce diversity in the ensemble models. As shown in our results, this technique can also push the predictions of models toward each other, and generates differences mostly at the boundary or within hard-to-segment regions. As future work, it would be interesting to explore a broader range of strategies to create diversity, for example using fake images from generative adversarial networks. Moreover, our experiments revolved around three different medical image segmentation problems and included images from both MRI and CT modalities. As motivated in the introduction, semi-supervised learning is most important for medical applications, where annotating images is complex and expensive. Nonetheless, evaluating the proposed method on additional types of images and segmentation tasks would help to further validate its usefulness.

CHAPTER 4

BOOSTING SEMI-SUPERVISED IMAGE SEGMENTATION WITH GLOBAL AND LOCAL MUTUAL INFORMATION REGULARIZATION

Jizong Peng[1] , Marco Pedersoli[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] Department of Automated Production, École de Technologie Supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3

Article published in Journal *«Machine Learning for Biomedical Imaging»* in June 2021

4.1 Presentation

The chapter presents the article “Boosting Semi-supervised Image Segmentation with Global and Local Mutual Information Regularization" by Peng, Pedersoli, and Desrosiers published by the journal *Machine Learning for Biomedical Imaging* on 8 June 2021. The initial results of this paper were accepted as a poster paper at *Medical Imaging with Deep Learning* conference 2020. In this paper, we present a novel semi-supervised segmentation method that leverages mutual information (MI) on categorical distributions to achieve both global representation invariance and local smoothness. We evaluate the method on four challenging publicly-available datasets for medical image segmentation. Experimental results show our method to outperform recently-proposed approaches for semi-supervised segmentation and provide an accuracy near to full supervision while training with very few annotated images

4.2 Introduction

Supervised learning approaches based on deep convolutional neural networks (CNNs) have achieved outstanding performance in a wide range of segmentation tasks. However, such approaches typically require a large amount of labeled images for training. In medical imaging applications, obtaining this labeled data is often expensive since annotations must be made by trained clinicians, typically in 3D volumes, and regions to segment can have very low contrast. Semi-supervised learning is a paradigm which reduces the need for fully-annotated data by exploiting the abundance of unlabeled data, i.e. data without expert-annotated ground truth. In contrast to standard approaches that learn exclusively from labeled data, semi-supervised methods also leverage intrinsic properties of unlabeled data (or *priors*) to guide the learning process.

Among the main approaches for semi-supervised segmentation, those employing consistency-based regularization and unsupervised representation learning have shown a great potential at exploiting unlabeled data (Perone & Cohen-Adad, 2018; Perone *et al.*, 2019; Bortsova *et al.*, 2019; Li, Yu, Chen, Fu & Heng, 2018b; Chaitanya *et al.*, 2020). The former approach, which leverages the principle of transformation equivariance, i.e., $f(T(x)) = T(f(x))$ for a geometrical transformation T, enforces the segmentation network to predict similar outputs for different transformed versions of the same unlabeled image (Perone & Cohen-Adad, 2018; Bortsova *et al.*, 2019; Li *et al.*, 2018b). Typical geometrical transformations include small translations, rotations or scaling operations on the image. A common limitation for consistency-based methods, however, is that they ignore the dense and structured nature of image segmentation, and impose consistency on different pixels independently. On the other hand, representation learning (Bengio, Courville & Vincent, 2013) uses unlabeled data in a pre-training step to find an internal representation of images (i.e., convolutional feature maps) which is useful to the downstream analysis task. A recent technique based on this paradigm is contrastive learning (Oord *et al.*, 2018; Tian, Krishnan & Isola, 2019). In this technique, a network is trained with a set of paired samples from the same joint distribution (positive pair) or different distributions (negative pair). A contrastive loss is employed to make the representation of positive-pair images similar to each other, and the representation of negative-pair images to be different. Despite showing encouraging results for segmentation (Chaitanya *et al.*, 2020), contrastive learning methods typically suffer from major drawbacks. In particular, they require a large number of negative pairs and a large batch size to work properly (Chen, Kornblith, Norouzi & Hinton, 2020a), which makes training computationally expensive for medical image segmentation. These drawbacks are primarily due to the use of a continuous-variable representation that makes the estimation of the joint distribution of samples or their mutual information more difficult (Poole, Ozair, Oord, Alemi & Tucker, 2019; Ji *et al.*, 2019).

An alternative approach to unsupervised representation learning, based on a discrete representation, is clustering (Ji *et al.*, 2019; Caron, Bojanowski, Joulin & Douze, 2018; Peng, Desrosiers & Pedersoli, 2019). In deep clustering, a network is trained with unlabeled data to map examples with similar semantic meaning to the same cluster label. The challenge of this unsupervised task is twofold. Firstly, using traditional pairwise similarity losses like KL divergence or L_2 leads to the trivial solution where all examples are mapped to the same cluster (Bridle, Heading & MacKay, 1992; Krause, Perona & Gomes, 2010; Hu, Miyato, Tokui, Matsumoto & Sugiyama, 2017b; Ji *et al.*, 2019). Also, unlike for supervised classification, the labels in clustering are arbitrary and any permutation of these labels gives an equivalent solution. To address these challenges, Ji *et al.* (2019) recently proposed an Information Invariant Clustering (IIC) algorithm based on mutual information (MI). The MI between two variables X and Y corresponds to the KL divergence between their joint distribution and the product of their marginal distributions:

$$I(X;Y) \;=\; D_{\mathrm{KL}}(p(X,Y)\,||\,p(X)\,p(Y)). \tag{4.1}$$

Alternatively, MI can also be defined as the difference between the entropy of Y and its entropy conditioned on X:

$$\begin{aligned} I(X;Y) \;&= H(Y) \,-\, H(Y|X) \\ &= \mathbb{E}_Y\,[\,\log \mathbb{E}_X[\,p(Y|X)\,]\,] \,-\, \mathbb{E}_{X,Y}[\,\log p(Y|X)\,]. \end{aligned} \tag{4.2}$$

The IIC algorithm seeks network parameters which maximize the MI between the cluster labels of different transformed versions of an image. As can be seen from Eq. (4.2), if X is a random variable corresponding to an image and Y is another variable representing a cluster label, this approach avoids the trivial assignment of all images the same cluster since the first term (entropy) is maximized for uniformly distributed clusters Y (Hu *et al.*, 2017b; Zhao *et al.*, 2019b).

Recently, Peng, Pedersoli & Desrosiers (2020b) adapted the IIC algorithm to semi-supervised segmentation. In their work, a network is trained with both labeled and unlabeled data such that its prediction for labeled images is similar to the ground-truth mask, and output labels for neighbor patches in different transformed versions of the same unlabeled image (after reversing the transform) have a high MI. This MI-based approach has two positive effects on segmentation. First, it makes the network more robust to image variability corresponding to the chosen transformations. Second, it increases the local smoothness of the segmentation and avoids collapse to a single class. Since MI is invariant to the permutation of cluster labels, another loss based on KL is also added to align these labels across different image patches during training. Although leading to improved performance for the various segmentation tasks, this recent method has the following two limitations: 1) it only regularizes the output of the network, not its internal representation; 2) the regularization is only applied locally in the image, and not globally.

4.2.1 Contributions

In this paper, we propose a novel semi-supervised segmentation method which uses the MI between representations computed at different hierarchical levels of the network to regularize its prediction both globally and locally. The proposed method employs auxiliary projection heads on layers of both the encoder and the decoder to group together feature vectors that are semantically related. Two separate strategies are used to achieve global and local regularization. In the global regularization strategy, we consider the entire feature map at a given layer as a representation of the input image and learn a mapping from this representation to a set of cluster labels. By maximizing the MI between the cluster assignments of two transformed versions of the same image, we thus promote invariance (equivariance) of the network with respect to the considered transformations. On the other hand, the local regularization strategy learns clusters for each spatial location of feature maps in the decoder, and maximizes the MI between cluster assignments of two neighbor feature vectors in transformed images. This enhances the spatial consistency of the segmentation output.

The detailed contributions are as follows:

- We propose the first semi-supervised segmentation method using MI maximization on categorical labels to achieve both global representation invariance and local smoothness. Our method is orthogonal to state-of-the-art consistency-based approaches like Mean Teacher which impose consistency only on the output space. By clustering feature embeddings from different hierarchical levels and scales, our method can effectively achieve a higher performance with very few labeled images.
- This paper represents a major extension of our previous work in (Peng *et al.*, 2020b) where clustering-based MI regularization was only applied locally on the network output. In contrast, the method proposed in this paper maximizes MI between both local and global feature embeddings from different layers of the network encoder and decoder. In a comprehensive set of experiments, we show that feature representations from separate hierarchical levels capture complementary information and contribute differently to performance. Moreover, we visually demonstrate the clustering effect of the proposed loss that maximizes MI between categorical labels.

The rest of this paper is as follows. In the next section, we give a summary of related work on semi-supervised segmentation and unsupervised representation learning. In Section 4.4, we then present the proposed semi-supervised segmentation method and explain how MI between cluster assignment labels is leveraged to achieve both local and global segmentation consistency. Our comprehensive experimental setup, involving four challenging segmentation datasets and comparing against strong baselines, is detailed in Section 5. Results, reported in Section 4.6, show our method to significantly outperform compared approaches and yield performance near to full supervision when trained with only 5% of labeled examples.

4.3 Related works

4.3.1 Semi-supervised segmentation

Although initially developed for classification (Oliver *et al.*, 2018), a wide range of semi-supervised methods have also been proposed for semantic segmentation. These methods are based on various learning techniques, including self-training (Bai *et al.*, 2017), distillation (Radosavovic *et al.*, 2018), attention learning (Min & Chen, 2018), adversarial learning (Souly *et al.*, 2017; Zhang *et al.*, 2017c), entropy minimization (Vu *et al.*, 2019), co-training (Peng, Estrada, Pedersoli & Desrosiers, 2020a; Zhou *et al.*, 2019b), temporal ensembling (Perone & Cohen-Adad, 2018), manifold learning (Baur *et al.*, 2017), and data augmentation (Chaitanya *et al.*, 2019; Zhao *et al.*, 2019a). Among recently proposed methods, consistency-based regularization has emerged as an effective way to improve performance by enforcing the network to output similar predictions for unlabeled images under different transformations (Bortsova *et al.*, 2019). Following this line of research, the Π model perturbs an input image with stochastic transformations or Gaussian noise and improves the generalization of a network by minimizing the discrepancy of its output for perturbed images. Virtual adversarial training (VAT) replaces the random perturbation with an adversarial one targeted at fooling the trained model. By doing so, the network efficiently learns a local smoothness prior and becomes more resilient to various noises. Consistency has also been a key component in temporal ensembling techniques like Mean Teacher (Perone & Cohen-Adad, 2018), where the output of a student network at different training iterations is made similar to that of a teacher network whose parameters are an exponential weighted temporal average of the student's. This method has shown great success for various semi-supervised tasks such as brain lesion segmentation (Cui *et al.*, 2019), spinal cord gray matter segmentation (Perone & Cohen-Adad, 2018) and left atrium segmentation (Yu *et al.*, 2019).

Despite improving performance in semi-supervised settings, a common limitation of the above methods is that they consider the prediction for different pixels as independent and apply a pixel-wise distance loss such as KL divergence or L_2 loss. This ignores the dense structure nature of the segmentation. Moreover, those approaches only regularize the output of the network for perturbed inputs, ignoring the hierarchical and multi-scale information found in different layers of the network.

4.3.2 Unsupervised representation learning

Important efforts have also been invested towards learning robust representations from unlabeled data. In self-supervised learning (Noroozi & Favaro, 2016; Kim, Cho, Yoo & Kweon, 2018; Noroozi, Vinjimoor, Favaro & Pirsiavash, 2018), unlabeled data are typically exploited in a first step to learn a given pretext task. This pretext task helps the network capture meaningful representations that can improve learning downstream tasks like classification or segmentation with few labeled data. Taleb, Lippert, Klein & Nabi (2019) trained a convolutional network to solve jigsaw puzzles and used the learned representation to boost performance for multi-modal medical segmentation. Other pretext jobs include predicting the transformation applied to an input image (Zhang, Qi, Wang & Luo, 2019; Wang, Kihara, Luo & Qi, 2019b) and converting a grey-scale image to RGB (Zhang, Isola & Efros, 2016).

Recently, contrastive learning was shown to be an effective strategy for semi-supervised learning. In this approach, one trains a network with a set of paired examples, together with a critic function to tell whether a pair of examples comes from their joint distribution or not. In their Contrastive Predicted Coding (CPC) approach, Oord *et al.* (2018) use a contrastive loss to learn a representation which can be predicted with an autoregressive model. Tian *et al.* (2019) proposed a Contrastive Multiview Coding (CMC) method where the network must produce similar features for images of different modalities if they correspond to the same object. Chen *et al.* (2020a) instead learn to predict whether a pair of images comes from a same image under different data augmentations. So far, only a single work has investigated contrastive learning for medical image segmentation (Chaitanya *et al.*, 2020). In this work, a network is trained to distinguish whether a pair of 2D images comes from the same physical position of their corresponding 3D volumes or not. Although contrastive learning has been shown to be related to MI (Tian *et al.*, 2019), the approach of Chaitanya *et al.* (2020) differs significantly from our method. First, their approach uses a standard contrastive loss between continuous vectors that requires sampling a large number of negative pairs and is expensive for image segmentation. In contrast our method exploits the MI between categorical labels, which can be computed efficiently. Moreover, whereas they impose consistency between corresponding positions in two different feature maps, our method also enforces it between neighbor positions and for different image transformations. This adds local smoothness to the feature representations and helps generate a more plausible segmentation. Last, whereas their approach only leverages unlabeled data in a pre-training step, we optimize the segmentation network with both labeled and unlabeled images in a single step.

Deep clustering has also been explored to learn robust representation of image data. Since it favors balanced clusters, thus avoiding the collapse of the solution to a single cluster, and does not make any assumption about the data distribution, MI has been at the core of several deep clustering methods. One of them, Information Maximizing Self-Augmented Training (IMSAT) (Hu *et al.*, 2017b), maximizes the MI between input data X and the cluster assignment Y. The output is regularized through the use of virtual adversarial samples (Miyato *et al.*, 2018), imposing that the original sample and the adversarial one should have a similar cluster assignment probability distribution. A related approach, called Invariant Information Clustering (IIC) (Ji *et al.*, 2019), instead maximizes the MI between cluster assignments of a sample and its transformed versions. Recently, Peng *et al.* (2020b) proposed an semi-supervised segmentation method inspired by IIC which encourages nearby patches in the network's output map, for two transformed versions of the same unlabeled image, to have a high MI. As mentioned above, this avoid the trivial assignment of all pixels to a single class and also promotes spatial smoothness in the segmentation. However, a common limitation of deep clustering methods for image classification and segmentation is that they only consider the network output, and ignore the rich semantic information of features inside the network.

4.3.3 Estimating MI

Capturing MI between two random variables is a difficult task, especially when these variables are continuous and/or high-dimensional. Traditional density- or kNN-based methods (Suzuki, Sugiyama, Sese & Kanamori, 2008; Vejmelka & Hlaváčková-Schindler, 2007) do not scale well to complex data such as raw images. Recently, variational approaches have become popular for estimating MI between latent representations and observations (Hjelm *et al.*, 2018; Oord *et al.*, 2018) or between two related latent representations (Tian *et al.*, 2019; Chaitanya *et al.*, 2020). These approaches instead maximize a variational lower bound to MI, thus making the problem tractable. Related to our work, Belghazi *et al.* (2018) leveraged the dual representation of KL divergence to develop a variational neural MI estimator (MINE) for image classification. In their Deep InfoMax method, Hjelm *et al.* (2018) used MINE to measure and maximize the MI between global and local representations. Various improvements have later been proposed to mitigate the high estimation variance of MINE (McAllester & Stratos, 2020), such as using f-divergence representation (Nowozin, Cseke & Tomioka, 2016), Jensen–Shannon (JS) divergence based optimization (Hjelm *et al.*, 2018; Zhao, Lu, Ma, Zhang & Zheng, 2020a), and clipping output with a prefixed range (Song & Ermon, 2019). Contrastive-based methods have been shown to underestimate MI (Hjelm *et al.*, 2018; McAllester & Stratos, 2020) and require a large number of negative examples (Tian *et al.*, 2019). As alternative to MINE, discriminator-based MI estimation (Liao, Moyer, Golland & Wells, 2020; Mukherjee, Asnani & Kannan, 2020) trains a *binary* classification network to directly emulate the density ratio between the joint distribution and the product of marginal.

Our method differs significantly from the above-mentioned approaches. First, these approaches usually define a *statistic network* (Belghazi *et al.*, 2018) or a discriminator (Liao *et al.*, 2020; Mukherjee *et al.*, 2020) to project high dimension data to a scalar, which often consists of convolution and MLP layers (Hjelm *et al.*, 2018; Liao *et al.*, 2020). On the contrary, our method employs a simple classifier to find proper categorical distributions and then maximize the estimated MI. This helps optimize the mutual information between dense representations efficiently. Compared to contrastive-based methods (Tian *et al.*, 2019; Chaitanya *et al.*, 2020), as we will show in Sec. 4.6.5, we can improve performance by simply increasing the number of clusters K instead of the batch size. The latter is not easily achieved in a memory- and computation-expensive task like segmentation. Last but not least, above-mentioned approaches rely on sampling *both* positive and negative pairs and seek to identify a *binary* decision boundary separating the joint distribution from the product of marginals. In contrast, we do not require negative pairs, similar to the recently proposed BYOL method (Grill *et al.*, 2020), but instead learn a fine-grain *multi-class* mapping. We leave as future work the comparison of different MI estimation strategies for semi-supervised segmentation.

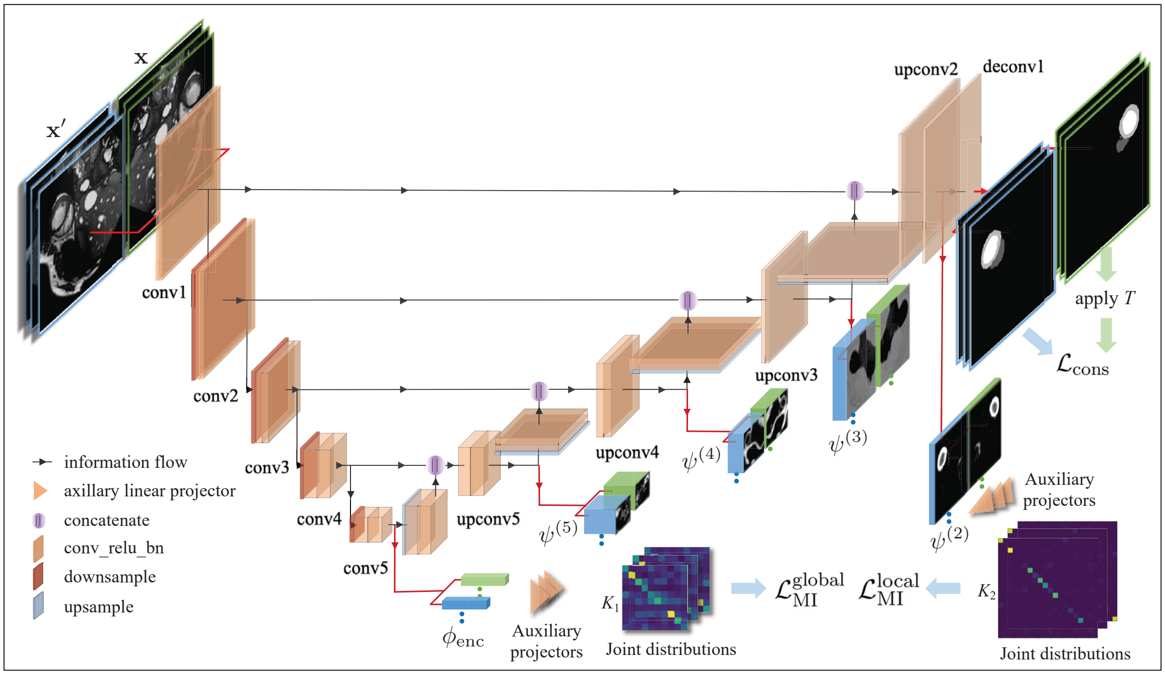


Figure 4.1 Training pipeline of our semi-supervised segmentation method. Given an unlabeled image x and its transformation x', we seek to maximize the mutual information of their intermediate feature representation with the help of auxiliary projectors. We maximize the global MI ($\mathcal{L}_{\text{MI}}^{\text{global}}$ loss) for embeddings taken from the encoder to learn transformation-invariant representation. Meanwhile, local MI is maximized ($\mathcal{L}_{\text{MI}}^{\text{local}}$ loss) for embeddings taken from the decoder, encouraging the network to group schematically-related regions while taking into consideration the spatial smoothness. $\mathcal{L}_{\text{cons}}$ further enforces the consistency on prediction distributions through different transformation and ensures the alignment of cluster label throughout the network.

4.4 Proposed method

We start by defining the problem of semi-supervised segmentation considered in this work and give and overview of the proposed method. We then explain each component of our method in greater details.

4.4.1 Semi-supervised segmentation model

We consider a semi-supervised segmentation task where we have a labeled dataset $\mathcal{D}_l$ of image-label pairs $(x, \mathbf{y})$, with image $x \in \mathbb{R}^{\Omega}$ and ground-truth labels $\mathbf{y} \in \{1, \ldots, C\}^{\Omega}$, and a larger unlabeled dataset $\mathcal{D}_u$ consisting of images without their annotations. Here, $\Omega = \{1, \ldots, W\} \times \{1, \ldots, H\}$ represents the image space (i.e., set of pixels) and C is the number of segmentation classes. We seek to learn a neural network f parametrized by θ to predict the segmentation label of each pixel of the input image.

Fig. 4.1 illustrates the proposed network architecture and training pipeline. We use an encoder-decoder architecture for the segmentation network, where encoder ϕ_{enc} extracts the information of an input image x by passing it through multiple convolutional blocks with down-sampling, and squeezes it into a compact embedding $\phi_{\text{enc}}(x)$. This embedding usually summarizes the global context of the image. The decoder ϕ_{dec} then gradually up-samples this embedding, possibly using some side information, and outputs the prediction $\mathbf{y} = \phi_{\text{dec}}(\phi_{\text{enc}}(x))$. While our method is agnostic to the choice of segmentation network, we consider in this work the well-known U-Net architecture (Ronneberger *et al.*, 2015) which achieved good performance on various bio-medical segmentation tasks. Compared to traditional encoder-decoder architectures, U-Net adds skip connections from the encoder to the decoder to reuse feature maps of same resolution in the decoder, thus helping to preserve fine details in the segmentation.

Following the main stream of semi-supervised segmentation approaches, our method exploits both labeled and unlabeled data during training. The parameters θ of the network are learned by optimizing the following loss function:

$$\mathcal{L}(\theta; \mathcal{D}_l, \mathcal{D}_u) \; = \; \mathcal{L}_{\text{spv}}(\theta; \mathcal{D}_l) + \lambda_1 \mathcal{L}_{\text{MI}}^{\text{global}}(\theta; \mathcal{D}_u) + \lambda_2 \mathcal{L}_{\text{MI}}^{\text{local}}(\theta; \mathcal{D}_u) + \lambda_3 \mathcal{L}_{\text{cons}}(\theta; \mathcal{D}_u). \quad (4.3)$$

This loss is comprised of four separate terms, which relate to different aspects of the segmentation and whose relative importance is controlled by hyper-parameters $\lambda_1, \lambda_2, \lambda_3 \geq 0$. As in standard supervised methods, $\mathcal{L}_{\text{spv}}$ uses labeled data $\mathcal{D}_l$ and imposes the pixel-wise prediction of the network for an annotated image to be similar to the ground truth labels. While other segmentation losses like the Dice loss could also be considered, our method uses the well-known cross-entropy loss:

$$\mathcal{L}_{\text{spv}}(\theta; \mathcal{D}_l) \;=\; -\frac{1}{|\mathcal{D}_l|\,|\Omega|} \sum_{(x,\mathbf{y})\in\mathcal{D}_l} \sum_{(i,j)\in\Omega} y_{ij} \log f_{ij}(x;\theta). \tag{4.4}$$

Since we have no annotations for images in $\mathcal{D}_u$, we instead use this unlabeled data to regularize the learning and guide the optimization process toward good solutions. This is achieved via three loss terms: $\mathcal{L}_{\text{MI}}^{\text{global}}$, $\mathcal{L}_{\text{MI}}^{\text{local}}$, and $\mathcal{L}_{\text{cons}}$. The first two are based on maximizing the MI between the feature embeddings of an image under different data augmentation, where embeddings can come from different hierarchical levels of both the encoder and the decoder. Specifically, we want to capture the information dependency between the semantically-related feature maps, while avoiding the complex computation of this dependency in continuous feature space. To obtain an accurate and efficient estimation of MI, we resort to a set of auxiliary projectors that convert features into categorical distributions.

We exploit this idea in two complementary regularization losses, focusing on global MI and local MI. The global MI loss $\mathcal{L}_{\text{MI}}^{\text{global}}$ considers the embedding $\phi_{\text{enc}}(x)$ produced by the encoder as a global representation of an image x, and enforces this representation to preserve its information content under a given set of image transformations. On the other hand, the local MI loss $\mathcal{L}_{\text{MI}}^{\text{local}}$ is based on the principle that information within a small region of the image should be locally invariant. That is, the MI between a vector in a feature map and its neighbor vectors should be high, if they correspond to the same semantic region of the image. By maximizing the MI between neighbor vectors, we can thus obtain feature representations and a segmentation output which are spatially consistent.

The last term in (4.3), $\mathcal{L}_{\text{cons}}$, is a standard transformation consistency regularizer that is included for two main reasons. First, as in regular consistency-based methods, it forces the network to produce the same pixel-wise output for different transformations of a given image, after reversing the transformations. Therefore, it directly promotes equivariance in the network. The second reason stems from the fact that MI is permutation-invariant and, thus, any permutation of labels in two cluster assignments does not change their MI. Hence, $\mathcal{L}_{\text{cons}}$ helps align those labels across the network. We note several differences between $\mathcal{L}_{\text{cons}}$ and $\mathcal{L}_{\text{MI}}^{\text{local}}$. While $\mathcal{L}_{\text{cons}}$ is only employed at the network output, $\mathcal{L}_{\text{MI}}^{\text{local}}$ may also be used at different layers of the decoder. Moreover, because it imposes strict equality, $\mathcal{L}_{\text{cons}}$ can only be used between corresponding pixels in two images. In contrast, $\mathcal{L}_{\text{MI}}^{\text{local}}$ also considers information similarity between feature map or output locations that are not in perfect correspondence. In the following subsections, we present each of the three regularization loss terms individually.

4.4.2 Global mutual information loss

Let x be an image sampled from $\mathcal{D}_u$ and T an image transformation drawn from a transformation pool $\mathcal{T}$. Transformation T is typically a random crop, horizontal flip, small rotation, or a combination of these operations. After applying T on x, the transformed image $x' = T(x)$ should share similar contextual information as x. Consequently, we expect a high MI between random variables corresponding to original and transformed images. Based on this idea, we want the encoder ϕ_{enc} to learn latent representations for these images which maximizes their mutual information:

$$\max_{\theta_{\text{enc}}} \; I(\phi_{\text{enc}}(X);\, \phi_{\text{enc}}(X')) \tag{4.5}$$

where θ_{enc} are the enconder's learnable parameters. However, optimizing directly Eq. (4.5) is notoriously difficult as the two variables are in continuous space. For instance, one has to learn a critic function and maximize a variational lower bound of I, which may result in heavy computation and high variance (Liao *et al.*, 2020; Song & Ermon, 2019).

To overcome this problem, we adapt the method proposed for unsupervised clustering and project the embeddings into categorical distributions $p(Z \,|\, x) = g(\phi_{\text{enc}}(x))) \in [0, 1]^K$ with an auxiliary projector g consisting of a linear layer followed by a softmax activation. Using this approach, embeddings $\phi_{\text{enc}}(x)$ and $\phi_{\text{enc}}(x')$ are converted to cluster probability distributions $p(Z \,|\, x)$ and $p(Z \,|\, x')$ with a predefined cluster number K. This projection introduces a bottleneck effect on (4.5) since

$$I(g(\phi_{\text{enc}}(X));\, g(\phi_{\text{enc}}(X'))) \;\leq\; I(\phi_{\text{enc}}(X);\, \phi_{\text{enc}}(X')) \tag{4.6}$$

The information bottleneck theory states that a capacity-limited network g can lead to information loss which results in a reduced MI between the two variables (Tishby, Pereira & Bialek, 2000; Alemi, Fischer, Dillon & Murphy, 2016; Ji *et al.*, 2019). The equality holds when g is an invertible mapping between embedding space to K categories, which is not the case for a linear projection g.

The conditional joint distribution of cluster labels

$$p(Z, Z' \,|\, x, x') \;=\; g(\phi_{\text{enc}}(x)) \cdot g(\phi_{\text{enc}}(T(x)))^\top \tag{4.7}$$

yields a $K \times K$ probability matrix for each $x \in \mathcal{D}_u$, $x' = T(x)$, and T sampled from $\mathcal{T}$. After marginalizing over the entire $\mathcal{D}_u$ (or a large mini-batch in practice), the $K \times K$ joint probability distribution $P = p(Z, Z')$ can be estimated as

$$P \;\approx\; \frac{1}{|\mathcal{D}_u|\,|\mathcal{T}|} \sum_{x \in \mathcal{D}_u} \sum_{T \in \mathcal{T}} g(\phi_{\text{enc}}(x)) \cdot g(\phi_{\text{enc}}(T(x)))^\top. \tag{4.8}$$

Using the definition of MI in (4.1), the proposed global MI loss can then computed from P as follows:

$$\mathcal{L}_{\text{MI}}^{\text{global}}(\theta; \mathcal{D}_u) \;=\; -I(P) \;=\; -\sum_{c=1}^{K} \sum_{c'=1}^{K} P_{c,c'} \,\log\, \frac{P_{c,c'}}{\sum_{c=1}^{K} P_{c,c'} \cdot \sum_{c'=1}^{K} P_{c,c'}} \tag{4.9}$$

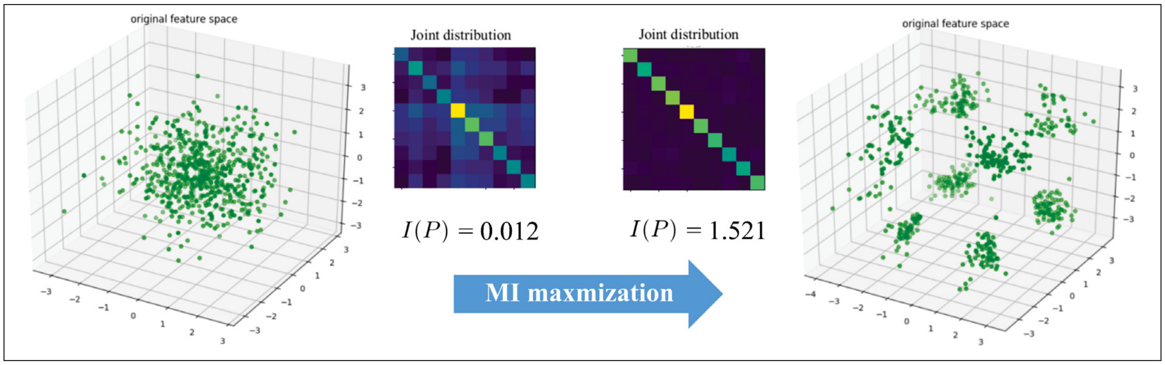


Figure 4.2 A toy example illustrating the effect of maximizing global MI. By increasing $I(P)$, randomly generated 3-D embedding points are effectively grouped into well-defined clusters.

A high $I(P)$ means that the information of $T(x)$ can be retrieved given x (i.e., low conditional entropy), which forces the encoder to learn transformation-invariant features and, more importantly, group together images with similar feature representations.

To illustrate the clustering effect of $\mathcal{L}_{\text{MI}}^{\text{global}}$, a simple example in 3-D feature space is presented in Fig. 4.2, where each randomly-generated point can be regarded as the three-dimensional embedding of an image. Optimizing $\mathcal{L}_{\text{MI}}^{\text{global}}$ groups these embeddings into multiple clusters based on their relative positions. As a result, the joint distribution P becomes confident with near-uniform values on the diagonal. This indicates that balanced clusters are formed and embedding points are pushed away from the decision hyperplanes defined by $g(\cdot)$.

4.4.3 Local mutual information loss

Our global MI loss focuses on the discriminative nature of encoder features, assuming that image-level contextual information can be captured. This may not be true for representations produced by decoder blocks. Given the features generated by the encoder, decoder blocks try to recover the spatial resolution of features and produce densely-structured representations. Therefore, features from the decoder will also capture local patterns that determine the final segmentation output. Based on this idea, we propose a local MI loss $\mathcal{L}_{\mathrm{MI}}^{\mathrm{local}}$ that preserves the local information of feature embeddings in the decoder.

Let $\psi^{(b)}(x) = \phi_{\mathrm{dec}}^{(b)}(\phi_{\mathrm{enc}}(x)) \in \mathbb{R}^{C_b \times H_b \times W_b}$ be the feature map produced in the b-th decoder block for an unlabeled image x. As described in Section 4.5.2, each block is composed of a convolution and an upsampling operation. This feature map has a reduced spatial resolution compared to x and segmentation output $\mathbf{y}$, and each of its feature vectors is a compact summary of a sub-region in the input image determined by the network's receptive field. Inspired by the fact that a region in an image shares information with adjacent, semantically-related ones, we maximize the MI between spatially-close elements of $\psi^{(b)}(x)$. Denoting as $[\psi^{(b)}(x)]_{i,j} \in \mathbb{R}^C$ the feature vector located at position (i, j) of the feature map, we define the neighbors of this vector using a set of displacement vectors $\Delta^{(b)} \subset \mathbb{Z}^2$:

$$\mathcal{N}_{i,j}^{(b)} \;=\; \{[\psi^{(b)}(x)]_{i+p,j+q} \,|\, (p,q) \in \Delta^{(b)}\}. \tag{4.10}$$

Furthermore, to make the decoder transformation invariant, we also enforce feature embeddings to have a high MI if they come from the same image under a data transformation $T \in \mathcal{T}$. Note that unlike for the global MI loss, where the feature map is considered as a single representation vector, we now have to align the two embeddings in a same coordinate system. Hence, we need to compare $[\psi^{(b)}(T(x))]_{i,j}$ with $[T(\psi^{(b)}(x))]_{i+p,j+q}$, where (p,q) is a displacement in $\Delta^{(b)}$. As before, we use a linear projection head h to convert the feature map $\psi^{(b)}(x)$ to a cluster assignment $h(\psi^{(b)}(x)) \in [0,1]^{K_b \times H_b \times W_b}$. Since we want to preserve the spatial resolution of the feature map, h is defined as a 1×1 convolution followed by a softmax. Following (Peng *et al.*, 2020b; Ji *et al.*, 2019), we then compute a separate joint distribution $P_{p,q}^{(b)}$ for each displacement $(p,q) \in \Delta^{(b)}$:

$$P_{p,q}^{(b)} \;\approx\; \frac{1}{|\mathcal{D}_u|\,|\mathcal{T}|\,|\Omega|} \sum_{x \in \mathcal{D}_u} \sum_{T \in \mathcal{T}} \sum_{(i,j) \in \Omega} h([\psi^{(b)}(T(x))]_{i,j}) \cdot h([T(\psi^{(b)}(x))]_{i+p,j+q})^{\top} \tag{4.11}$$

Note that the operation in (4.11) can be computed efficiently with standard convolution operations. Finally, we obtain the local MI loss by averaging the MI over all decoder blocks $b \in \{1, \ldots, B\}$ and corresponding displacements:

$$\mathcal{L}_{\mathrm{MI}}^{\mathrm{local}}(\theta; \mathcal{D}_u) \;=\; -\frac{1}{B}\sum_{b=1}^{B}\frac{1}{|\Delta^{(b)}|}\sum_{(p,q)\in\Delta^{(b)}} I(P_{p,q}^{(b)}) \tag{4.12}$$

where $I(P_{p,q}^{(b)})$ is computed as in (4.9).

4.4.4 Consistency-based loss

As we will show in experiments, employing only the MI-based regularization losses may be insufficient to achieve optimal performance. This is in part due to the clustering nature of these losses: for two distributions conditionally independent given the same input image, MI is maximized if there is a deterministic mapping between clusters in each distribution such that they are equivalent. For example, permuting the cluster labels in one of the two distributions does not change their MI.

To ensure the alignment of cluster labels throughout the network, we add a final loss term $\mathcal{L}_{\mathrm{cons}}$ which imposes the network output at each pixel of an unlabeled image to remain the same under a set of transformations. In this work, we measure output consistency using the L_2 norm:

$$\mathcal{L}_{\mathrm{cons}}(\theta; \mathcal{D}_u) \;=\; \frac{1}{|\mathcal{D}_u|\,|\mathcal{T}|\,|\Omega|}\sum_{x\in\mathcal{D}_u}\sum_{T\in\mathcal{T}}\sum_{(i,j)\in\Omega} \|f_{ij}(T(x)) - T(f_{ij}(x))\|_2^2 \tag{4.13}$$

This loss, which is typical to approaches based on transformation consistency, has been shown to boost segmentation performance in a semi-supervised setting (Bortsova *et al.*, 2019).

4.5 Experimental setup

4.5.1 Dataset and metrics

To assess the performance of the proposed semi-supervised method, we carried out extensive experiments on four clinically-relevant benchmark datasets for medical image segmentation: the Automated Cardiac Diagnosis Challenge (ACDC) dataset (Bernard *et al.*, 2018), the Prostate MR Image Segmentation (PROMISE) 2012 Challenge dataset (Litjens *et al.*, 2014), the Spleen sub-task dataset of the Medical Segmentation Decathlon Challenge (Simpson *et al.*, 2019), and the Multi-Modality Whole Heart Segmentation (MMWHS) dataset (Zhuang & Shen, 2016). These four datasets contain different image modalities (CT and MRI) and acquisition resolutions.

4.5.1.0.1 ACDC dataset

The publicly-available ACDC dataset consists of 200 short-axis cine-MRI scans from 100 patients, evenly distributed in 5 subgroups: normal, myocardial infarction, dilated cardiomyopathy, hypertrophic cardiomyopathy, and abnormal right ventricles. 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. Segmentation masks delineate 4 regions of interest: left ventricle endocardium (LV), left ventricle myocardium (Myo), right ventricle endocardium (RV), and background. We consider the 3D-MRI scans as 2D images through-plane due to the high anisotropic acquisition resolution, and re-sample them to a fix space ranging of 1.0×1.0 mm. Pixel intensities are normalized based on the 1% and 99% percentile of the intensity histogram for each patient. Normalized slices are then cropped to 384×384 pixels to slightly adjust the foreground delineation of the ground truth. For the main experiments, we used a random split of 8 fully-annotated and 167 unlabeled scans for training, and the remaining 25 scans for validation. In another experiment, we also evaluate our model trained with a varying number of patient scans as labeled data. A rich set of data augmentation was employed for both labeled and unlabeled images, including random crops of 224×224 pixels, random flip, random rotation within $[-45, 45]$ degrees, and color jitter.

4.5.1.0.2 Prostate dataset

This second dataset focuses on prostate segmentation and is composed of multi-centric transversal T2-weighted MR images from 50 subjects. These 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 $15 \times 256 \times 256$ to $54 \times 512 \times 512$ voxels with a spacing ranging from $2 \times 0.27 \times 0.27$ to $4 \times 0.75 \times 0.75$ mm^3. 2D images are sliced along short-axis and are resized to a resolution of 256×256 pixels. A normalization is then applied on pixel intensity based on 1% and 99% percentile of the intensity histogram for each patient. We randomly selected 4 patients as labeled data, 36 as unlabeled data, and 10 for validation during the experiments. For data augmentation, we employ the same set of transformation as the ACDC dataset, except we limit the random rotation to $[-10, 10]$ degrees.

4.5.1.0.3 Spleen dataset

The third dataset consists of patients undergoing chemotherapy treatment for liver metastases. A total of 41 portal venous phase CT scans were included in the dataset with acquisition and reconstruction parameters described in (Simpson *et al.*, 2019). The ground truth segmentation was generated by a semi-automatic segmentation software and then refined by an expert abdominal radiologist. Similar to the previous dataset, 2D slices are obtained by slicing the high-resolution CT volumes along the axial plane. Each slice is then resized to a resolution of 512×512 pixels for the sake of normalization. To evaluate algorithms in a semi-supervised setting, we randomly split the dataset into labeled, unlabeled and validation image subsets, comprising CT scans of 6, 30, and 5 patients respectively. For data augmentation, we employ a random crop of 256×256 pixels, color jitter, random horizontal flip, and random rotation of $[-10, 10]$ degrees.

4.5.1.0.4 Multi-Modality Whole Heart Segmentation (MMWHS) dataset

The last dataset includes 20 high-resolution CT volumes from 20 patients. The in-plane resolution is around 0.78×0.78mm and the average slice thickness is 1.60 mm. Following the same protocol as for the ACDC dataset, we prepossessed and sliced three dimensional images into 2D slices with a fixed space ranging of 1.0×1.0 mm. All slices were then center-cropped to 256×256 pixel. We randomly split the dataset into labeled (2 patients), unlabeled (13 patients) and validation (5 patients) sets, which were fixed throughout all experiments. We employ the same set of data augmentations as for ACDC.

For all the datasets, we used the commonly-adopted Dice similarity coefficient (DSC) metric to evaluate segmentation quality. DSC measures the overlap between the predicted labels (S) and the corresponding ground truth labels (G):

$$\mathrm{DSC}(S, G) = \frac{2|S \cap G|}{|S| + |G|} \tag{4.14}$$

DSC values range between 0 and 1, a higher value corresponding to a better segmentation. In all experiments, we reconstruct the 3D segmentation for each patient by aggregating the predictions made for 2D slices and report the 3D DSC metric for the validation set.

4.5.2 Implementation details

4.5.2.0.1 Network and parameters

For all four datasets, we employ the same U-Net architecture comprised of 5 Convolution + Downsampling blocks in the encoder, 5 Convolution + Upsampling blocks in the decoder, and skip connections between convolutional blocks of same resolution in the encoder and decoder (see Fig. 4.1 for details). We adopted this architecture as it was shown to work well for different medical image segmentation tasks.

Network parameters are optimized using stochastic gradient descent (SGD) with the Adam optimizer. For all experiments, we applied a learning rate warm-up strategy to increase the initial learning rate of 1×10^{-7} for both ACDC and MMWHS, 1×10^{-6} for Prostate and 1×10^{-6} for Spleen by a factor of 400 in the first 10 epochs and decreases it with a cosine scheduler for the following 90 epochs. We define an epoch as 300 iterations, each consisting of a batch of 4 labeled and 10 unlabeled images drawn with replacement from their respective dataset. The proposed MI-based regularization is applied to the feature embeddings generated in three different blocks: the last block of the decoder (**Conv5**) for the global MI loss, and the last two convolutional blocks from the decoder (**Upconv3** and **Upconv2**) for the local MI loss. In an ablation study, we measure the contribution of regularizing each of these embeddings on segmentation performance.

We employ an array of five linear projectors, instead of a single projector, to project feature embeddings to a corresponding set of categorical distributions, and average the MI-based losses over these distributions. For the encoder, the projector head consists of a max-pooling layer to summarize context information, a linear layer and a softmax activation layer. On the other hand, for the decoder, we only use a 1×1 convolution with softmax activation layer. As the proposed MI-based losses are computed on their output, the parameters of these projectors are also updated during training. We also tested projection head consisting of several layers with non-linearity, however this resulted in a similar performance but a higher variance. In the default setup of our method, we fixed the number of clusters to $K = 10$ for both the encoder and the decoder. In another ablation study, we show that a slightly greater performance can be achieved for a larger K, at the cost of increased computations.

To balance the different regularization terms in (4.3), we used weights of $\lambda_1 = 0.1$, $\lambda_2 = 0.1$ and $\lambda_3 = 5$ for experiments on the ACDC and MMWHS datasets, $\lambda_1 = 0.05$, $\lambda_2 = 0.05$ and $\lambda_3 = 10$ for the Prostate dataset, and $\lambda_1 = 0.05$, $\lambda_2 = 0.05$ and $\lambda_3 = 5$ for the Spleen dataset. These hyper-parameters were determined by grid search. We set the pool of transformations on unlabeled images ($\mathcal{T}$) as random horizontal and vertical flips. For the local MI loss, we set the neighborhood size Δ to be 3×3 for **Upconv3** and 7×7 for **Upconv2**, corresponding to a regions of 3-5 mm in original image space depending on the resolution. We also tested our method with larger neighborhoods, however this increased computational cost without significantly improving accuracy.

4.5.2.0.2 Compared methods

We compared our method against several baselines, ablation variants of our method and recently-proposed approaches for semi-supervised segmentation:

- **Full supervision**: We trained the network described above using the supervised loss $\mathcal{L}_{\text{spv}}$ on *all* training images. This results in an upper bound on performance.
- **Partial Supervision**: A lower bound on performance is also obtained by optimizing $\mathcal{L}_{\text{spv}}$ only on *labeled* images, ignoring the unlabeled ones.
- **Mutual information**: This ablation variant of our method consists in maximizing MI for intermediate feature embeddings while ignoring the consistency constraint on the output space (i.e., dropping $\mathcal{L}_{\text{cons}}$ in the loss).
- **Consistency regularization** (Bortsova *et al.*, 2019): This second ablation variant, which can be seen as the Π model for image segmentation, imposes $\mathcal{L}$cons loss as the only regularization loss, without using $\mathcal{L}_{\text{MI}}^{\text{global}}$ or $\mathcal{L}_{\text{MI}}^{\text{local}}$. Only the output distribution space is regularized while embeddings from intermediate features are unconstrained.
- **Entropy minimization** (Vu *et al.*, 2019): In addition to employing $\mathcal{L}_{\text{spv}}$ on labeled data, this well-known semi-supervised method minimizes the pixel-wise entropy loss of predictions

made for unlabeled images. By doing so, it forces the network to become more confident about its predictions for unlabeled images. To offer a fair comparison, we performed grid search on the hyper-parameter balancing the two loss terms, and report the score of the best found hyper-parameter.

- **Mean Teacher** (Perone *et al.*, 2019): This last approach 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. We use the formulation similar to (Perone *et al.*, 2019) and quantify the distribution difference using L_2 loss. Following the standard practice, we fix the decay coefficient to be 0.999. The coefficient balancing the supervised and regularization losses is once again selected by grid search.

All tested methods are implemented in a single framework, which can be found here: https://github.com/jizongFox/MI-based-Regularized-Semi-supervised-Segmentation.

Additional experiments To further assess the improvement on segmentation quality brought by the proposed global and local MI losses, we performed additional experiments on the ACDC dataset. We first compare our method against Mean Teacher for different amounts of labeled data. Second, we examine the sensitivity of our method to different regularization weights. Third, we investigate the importance of features from different hierarchical levels of the network. Fourth, we evaluate the impact on segmentation quality of using a different number of clusters K in projection heads of the proposed architecture. Finally, we show that the joint optimization of labeled and unlabeled data losses works better in our semi-supervised setting than a two-step strategy of pre-training the network on a clustering task using unlabeled data, and then fine-tuning it on the segmentation task with labeled data.

Table 4.1 Mean 3D DSC of tested methods on the ACDC, Prostate, Spleen and MMWHS datasets. RV, Myo and LV refer to the right ventricle, myocardium and right ventricle classes, respectively. Mutual information corresponds to our method without loss term $\mathcal{L}_{\text{cons}}$ and Consistency regularization corresponds to our proposed loss without $\mathcal{L}_{\text{MI}}^{\text{global}}$ or $\mathcal{L}_{\text{MI}}^{\text{local}}$. For ACDC, Prostate, Spleen and MMWHS, respectively 5%, 10%, 16.7% and 13.3% of training images are considered as annotated and the rest as unlabeled. Reported values are averages (standard deviation in parentheses) for 3 runs with different random seeds.

	ACDC						
	RV	**Myo**	**LV**	**Mean**	**Prostate**	**Spleen**	**MMWHS**
Full supervision	87.64 (0.46)	87.46 (0.15)	93.55 (0.33)	89.55 (0.29)	87.70 (0.13)	95.32 (0.70)	88.91 (0.12)
Partial Supervision	57.67 (1.54)	69.68 (2.35)	86.08 (1.15)	71.14 (1.28)	41.63 (2.41)	88.20 (1.89)	48.50 (1.73)
Entropy min.	56.69 (3.56)	73.46 (1.53)	86.80 (2.36)	72.32 (1.22)	55.47 (2.05)	90.77 (0.92)	49.44 (1.43)
Mean Teacher	80.04 (0.48)	81.81 (0.17)	90.44 (0.33)	84.10 (0.26)	80.61 (1.63)	93.12 (0.57)	55.57 (0.48)
Ours (MI only)	78.73 (0.82)	79.38 (0.40)	88.80 (0.66)	82.30 (0.57)	74.75 (1.89)	92.46 (0.80)	50.66 (1.38)
Ours (Consistency only)	75.21 (0.94)	82.31 (0.19)	**91.91 (0.47)**	83.14 (0.44)	77.92 (1.20)	94.19 (0.62)	49.15 (0.77)
Ours (all)	**81.87 (0.54)**	**83.65 (0.26)**	91.76 (0.32)	**85.76 (0.16)**	**81.76 (0.71)**	**94.61 (0.65)**	**55.75 (0.40)**

4.6 Experimental Results

4.6.1 Comparison with the state-of-the-art

Table 4.1 reports the mean 3D DSC obtained by tested methods on the validation set of the ACDC, Prostate, Spleen and MMWHS datasets. While using limited labeled data in training (e.g., 5% of the training set as labeled data for ACDC), large performance gaps are observed between partial and full supervision baselines, leaving space for improvements to the regularization techniques. Overall, all semi-supervised approaches tested in this experiment improved performance compared to the partial supervision baseline, showing the importance of also considering unlabeled data during training. Entropy minimization, the worse-performing semi-supervised baseline, yielded absolute DSC improvements of 1.20%, 13.83%, 2.57% and 0.94% for the ACDC, Prostate, Spleen and MMWHS datasets, respectively. Mean Teacher and Consistency regularization gave comparable results, both of them outperforming Entropy minimization by a large margin. This demonstrates the benefit of enforcing output consistency during learning, either directly or across different training iterations as in Mean Teacher. With respect to these strong baselines, the proposed method achieved a higher 3D DSC in all but one case (left ventricle segmentation in ACDC). When averaging performance over the RV, Myo and LV segmentation tasks of ACDC, our method obtains the highest mean DSC of 85.76%, compared to 84.10% for Mean Teacher and 83.14% for Consistency regularization. These improvements are statistically significant in a one-sided paired t-test ($p < 0.01$). The robustness of our method to the execution random seed (network parameter initialization, batch selection, etc.) can also be observed by the low standard deviation values obtained for all datasets and tasks.

The results in Table 4.1 show that the combination of the MI-based and consistency-based losses in the proposed method are essential to its success. Considering only MI maximization (Mutual Information method) yields a mean DSC improvement of 11.16% over the Partial Supervision baseline for ACDC, whereas performance is boosted by 12.00% when also enforcing transformation consistency on the output. Similar results are obtained for the Prostate and Spleen datasets. As mentioned before, this could be explained by the fact that MI is invariant to label permutation, therefore a pixel-wise consistency loss such as L_2 is necessary to align these labels across different cluster projections of features. The performance of our method can be appreciated visually in Fig. III-7, which shows examples of segmentation results for the tested methods. It can be seen that our method gives spatially-smoother segmentation contours that better fit those in the ground-truth. This results from regularizing network features both globally and locally. In contrast, only enforcing output consistency as in Mean Teacher and Consistency regularization leads to a noisier segmentation.

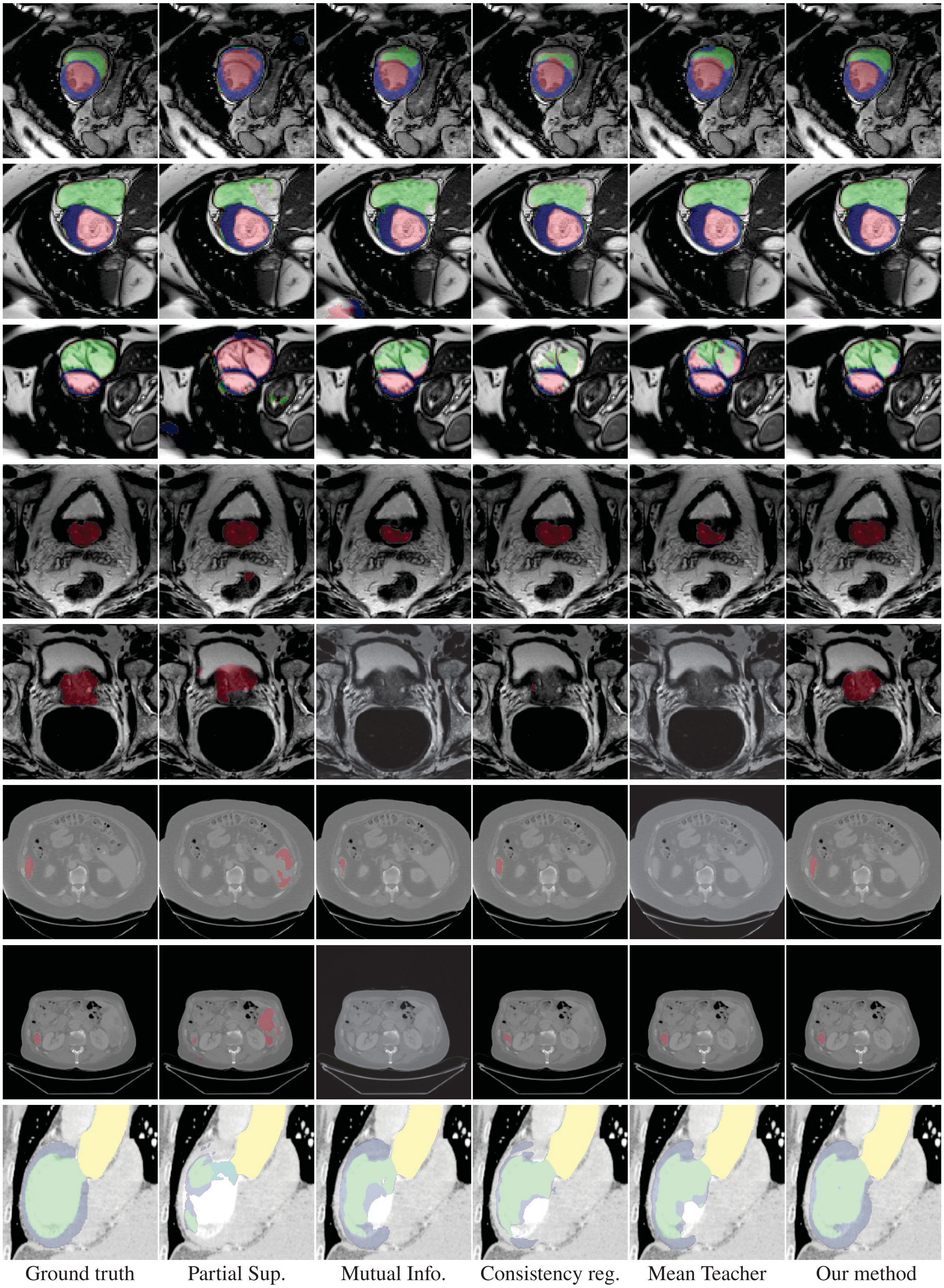


Figure 4.3 Visual comparison of tested methods on validation images. **Rows 1–3**: ACDC; **Rows 4–5**: Prostate; **Rows 6–7**: Spleen; **Row 8**: MMWHS.

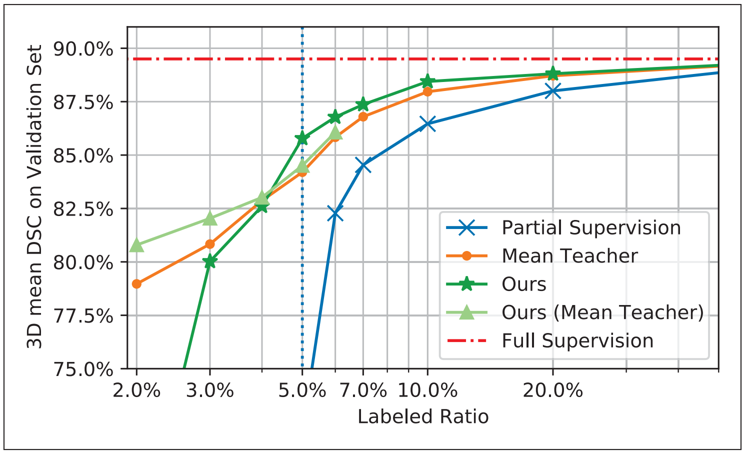


Figure 4.4 ACDC validation DSC versus various labeled data ratio for tested methods. It is clearly observed that our proposed method achieves higher performance compared with the state-of-the-art Mean Teacher in the regime of labeled ratio ≥ 5%. For an extreme case where only $2-3\%$ data are provided with annotations, our enhanced adaption outperforms Mean Teacher.

4.6.2 Impact of labeled data ratio

We further assess our method's ability to perform in a low labeled-data regime by training it with a varying number of labeled examples from the ACDC dataset, ranging from 2% to 50% of available training samples. As illustrated in Fig. 4.4, the proposed method (Dark green) offers a consistently better segmentation performance compared to Mean Teacher when over 5% of training examples are annotated. By exploiting temporal ensembling, Mean Teacher (Orange) provides a more plausible segmentation when given an extremely limited amount of labeled data (less than 3% of training samples). This can be attributed to the fact that, when trained with very limited labeled data, a single neural network is likely to overfit on those few examples and thus yield poor predictions for unlabeled images. Mean Teacher works well in this case as it exploits a separated Teacher network that distillates the knowledge of the student acquired at different training epochs, thereby implicitly smoothing the optimization and providing more a stable prediction on unlabeled images.

Table 4.2 ACDC validation DSC of our method using feature embeddings from different network layers. 5% of training samples are considered as labeled.

$\mathcal{L}_{\text{MI}}^{\text{global}}$	$\mathcal{L}_{\text{MI}}^{\text{local}}$				**ACDC validation DSC**				
Conv5	**Upconv5**	**Upconv4**	**Upconv3**	**Upconv2**	**RV**	**Myo**	**LV**	**Mean**	**Gain**
✓					76.29	81.83	90.65	82.92	11.78
	✓				76.34	82.61	91.68	83.54	12.40
		✓			80.04	81.39	90.00	83.81	12.67
			✓		**81.16**	**83.30**	**91.99**	**85.48**	**14.34**
				✓	77.65	81.96	90.90	83.50	12.36
✓	✓				78.03	82.86	91.31	84.07	12.93
✓		✓			76.63	81.33	90.54	82.83	11.69
✓			✓		**81.56**	**83.14**	**92.17**	**85.62**	**14.48**
✓				✓	77.54	82.00	90.05	83.20	12.06
✓	✓	✓			79.88	82.72	91.51	84.70	13.56
✓	✓		✓		78.24	82.94	91.67	84.28	13.14
✓	✓			✓	77.58	82.27	90.15	83.33	12.19
✓		✓	✓		79.80	82.21	90.90	84.30	13.16
✓		✓		✓	79.11	83.17	91.18	84.49	13.35
✓			✓	✓	**81.87**	**83.65**	**91.76**	**85.76**	**14.62**

Since the two methods are orthogonal, we can enhance our method by adapting it to the teacher-student framework of Mean Teacher. Toward this goal, we instead maximize the MI between feature embeddings of the teacher and the student, where the teacher's weights are computed as an expected moving average (EMA) of the student's. From Fig. 4.4 (Light green), we see that this enhanced version of our method offers a good trade-off between Mean Teacher and our default model. While its performance is similar to Mean Teacher for a labeled data ratio of 4% or more, it give a higher DSC when fewer annotated examples are provided. Thus, it improves the performance of Mean Teacher by 1.84% when only 2% of training samples are annotated.

4.6.3 Sensitivity to regularization loss weights

We carried out experiments on the ACDC dataset to investigate the relative impact on performance of the loss terms in Eq. (4.3), as defined by weights λ_1, λ_2 and λ_3. To simplify the analysis, the weights controlling our global and local mutual information losses are set to the same value $\lambda_1 = \lambda_2 = \lambda_{\text{MI}}$. The relative weight of the consistency loss, i.e. λ_3, is considered separately. We

Table 4.3 Mean DSC performance on the ACDC dataset given different λ_{MI} ($\lambda_{\text{MI}} = \lambda_1 = \lambda_2$) and λ_{con}. 5% of training samples are considered as labeled.

λ_{con}	$\lambda_{\text{MI}} = \lambda_1 = \lambda_2$				
	0.01	0.05	0.1	0.5	1.0
1.0	83.24%	85.3%	85.48%	83.79%	81.19%
5.0	84.46%	85.12%	**85.76%**	84.05%	82.30%
10.0	84.47%	85.41%	85.44%	83.59%	81.12%
15.0	84.18%	85.61%	85.17%	83.79%	80.89%

denote this weight as λ_{con} in the following results.

Table 4.3 reports the mean 3D DSC performance on the ACDC dataset using 5% of annotated data, for different combinations of λ_{MI} and λ_{con}. We see that λ_{MI} has a significant impact on segmentation performance. In general, DSC increases when λ_{MI} goes from 0.01 to 0.1, and decreases rapidly when for larger values. In contrast, our method is less sensitive to the choice of λ_{con}, indicating that the proposed global and local MI-based losses contribute most to segmentation quality in a semi-supervised setting.

4.6.4 Impact of embedding layers

The proposed MI-based losses regularize intermediate feature embeddings from both the encoder and decoder. The third experiment seeks to determine the impact of considering feature maps in different layers on results for the ACDC dataset. Since the global MI loss only uses features from the encoder's **Conv5** layer, we consider settings with and without this loss. On the other hand, the local MI loss regularizes features in four layers of the decoder: **Upconv2**-**Upconv5**. For our experiment, we test different combinations using a single or two of these layers in the local MI loss. Except for the selected features embeddings, the same training setting is used in all cases. Note that we set the neighborhood size Δ to be $1{\times}1$ for both **Upconv5** and **Upconv4** as their resolution scales correspond to $1/8$ and $1/4$ of an input image.

Table 4.4 Mean DSC performance on the ACDC dataset given different number of clusters K for the encoder and decoder. 5% of training samples are considered as labeled.

Decoder K	**Encoder** K			
	2	5	10	20
2	84.37%	84.66%	84.58%	84.56%
5	84.70%	85.43%	85.86%	86.05%
10	84.86%	85.33%	85.76%	85.91%
20	85.01%	85.52%	86.06%	**86.32%**

We observe from Table 4.2 that the choice of layers at which features are regularized has a noticeable impact results. Regularizing only encoder features (**Conv5**) in the global MI loss offers the smallest benefit. This may be due to the fact that segmentation requires learning the dense structure of an image, which is not well captured by the low-resolution features of the encoder. Conversely, highest improvements come from cases where **Upconv3** is selected in the local MI loss. The feature map in this layer has $1/2$ the resolution of the input image and, therefore, captures both global and local information. Overall, the best configuration is obtained with a combination of global regularization (**Conv5**) and local regularization (**Upconv3** and **Upconv2**), yielding a 14.62% gain in DSC over the Partial Supervision baseline.

4.6.5 Impact of cluster number K

A key component of our method is using auxiliary projectors to convert continuous feature representations to discrete cluster assignments. This encourages the grouping of semantically-related images/regions and enables the efficient computation of MI. As a result, the number of clusters K at each layer may also impact performance: if K is too small, image/region representations can only be grouped into a few discrete categories and, consequently, the network may fail to fully capture dependencies in the data. On the other hand, employing a very large K requires having a large batch size and can result in high variance.

In the next experiment, we tested different combinations of hyper-parameter $K \in \{2, 5, 10, 20\}$

for cluster assignments in the encoder (global MI loss) and decoder (local MI loss). Results of this experiments are summarized in Table 4.4. It can be seen that using a small $K = 2$ for the encoder and decoder results in relatively low performance. Furthermore, increasing the number of clusters in either or both parts of the network generally improves segmentation quality. However, employing a larger K also increases the computational cost of the method, especially for the local MI loss which relies on more expensive convolutional operations. On the whole, a value of $K = 10$ offers a good trade-off between segmentation performance and run-time complexity.

4.6.6 Visualization of clusters

Our method uses auxiliary projectors to map feature embeddings of corresponding images into categorical distributions. As mentioned before, this has a clustering effect where embeddings sharing similar semantic or structural information are grouped together while those with distinct information are pushed away. To illustrate this effect, we consider the ACDC dataset and plot in Fig. 4.5 the channel with highest activation at different positions of feature maps corresponding to decoder layers **Upconv3** and **Upconv2**. For visualization purposes, index values are mapped to the grey scale (min. index mapped to 0 and max. index to 255). The resulting channel map of our method is compared with those obtained using Partial Supervision and Mean Teacher. Moreover, we give in Fig. 4.6 the t-SNE plot of feature vectors at each position of the feature map in **Upconv2**, color-coded by the ACDC classes.

We observe in Fig. 4.5 that Partial Supervision outputs unrealistic predictions (rightmost column) and noisy feature activations (second and third columns). When trained with insufficient annotated data, a network can be misguided to learn noisy signals, such as local texture and geometric variability. In contrast, Mean Teacher and the proposed method produce segmentation maps similar to the ground truth. However, the feature activations of Mean Teacher appear noisier and less structured than those learned by our method. This confirms that regularizing

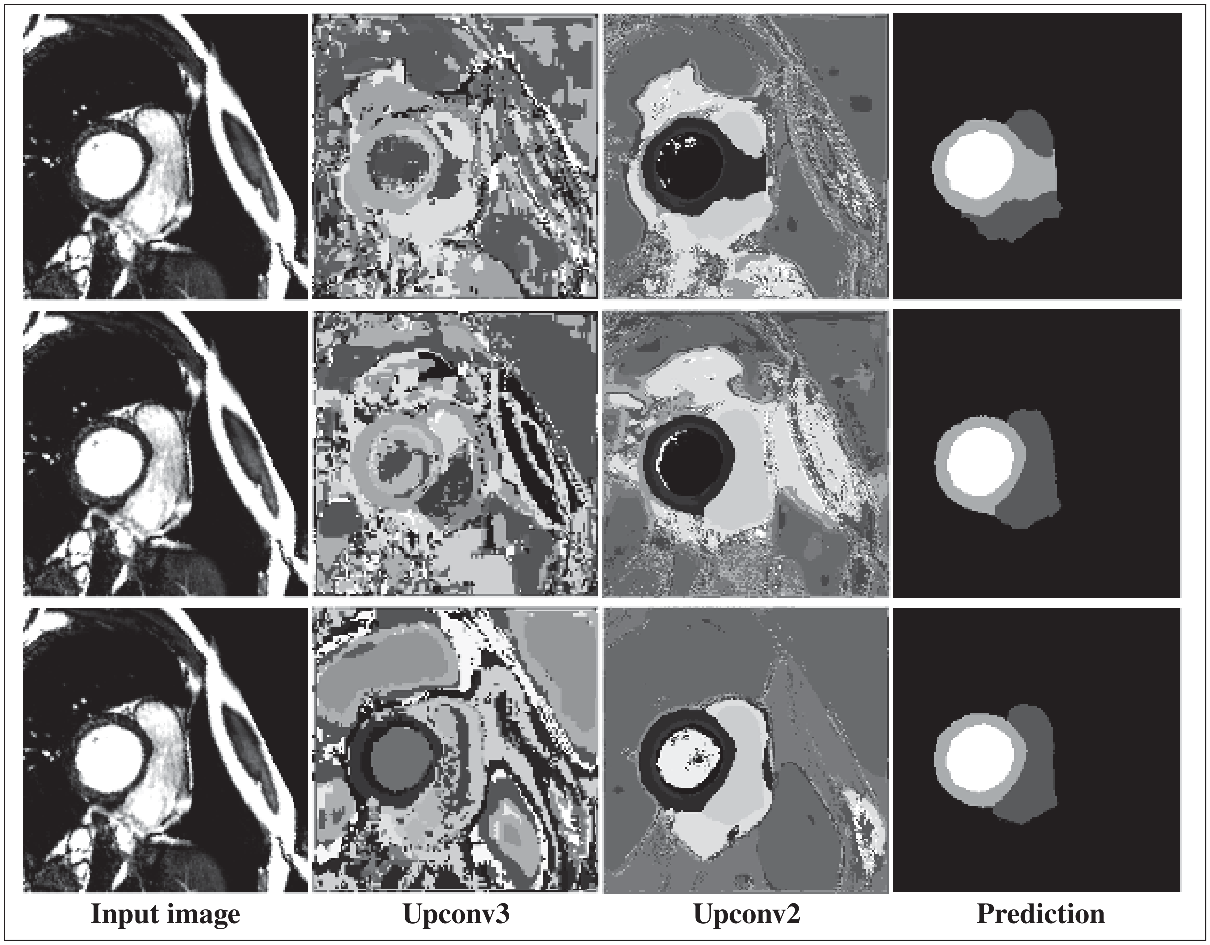


Figure 4.5 Visual comparison of maximum activations taken from network decoder positions. **Top row**: Partial Supervision. **Middle row**: Mean Teacher. **Bottom row**: Our method.

only the output space results in a poor internal representation. In comparison, the feature activations of our method better correlate with the semantic information of ground truth labels. This result confirmed by the 2D t-SNE plot in Fig. 4.6, where nearby points corresponds to positions in the feature map with similar feature vectors. As can be seen, our method exhibits more compact clusters with less outliers compared to Partial Supervision and Mean Teacher. This spatial clustering effect leads to a smoother segmentation and reduces overfitting when training with limited supervision.

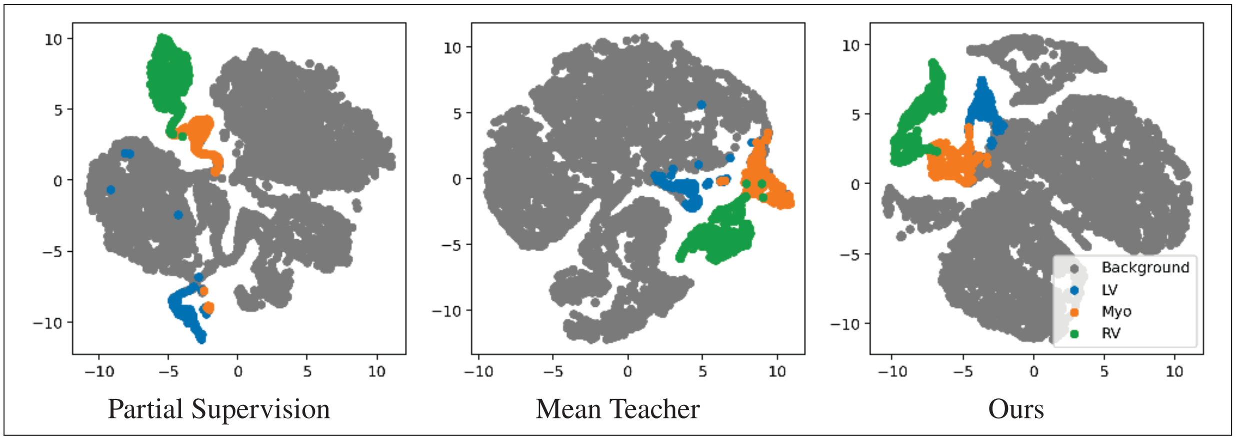


Figure 4.6 t-SNE plot on the ACDC validation set for different classes.

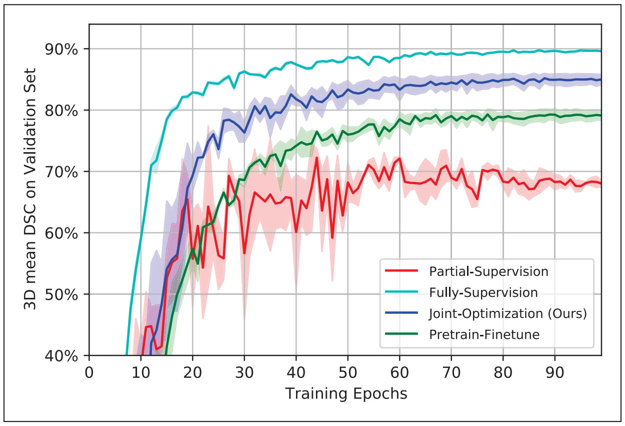


Figure 4.7 Validation mean DSC versus training epochs for the Pretrain-Finetune strategy and our Joint-Optimization method. Mean and standard deviation values are calculated from three independent runs.

4.6.7 Joint optimization of supervised and unsupervised losses

A significant difference between our method and (Ji *et al.*, 2019) relates to how models are trained. The approach in (Ji *et al.*, 2019) employs a two-stage training strategy where a feature representation is first obtained in an unsupervised way (clustering task), and then a mapping from clusters to segmentation labels is found based on a few labeled examples. In contrast, our method jointly optimizes both the supervised loss and unsupervised regularization terms during the whole training process.

The next experiment on the ACDC dataset is designed to validate the advantage of joint optimization compared to the two-stage training procedure, called Pretrain-Finetune in the following results. For this experiment, we use a similar setup as in previous experiments (e.g., 5% of training set as annotated examples) but make the following changes. For the first stage, we optimize a randomly-initialized segmentation network on *all* images using Eq. (4.3), without the supervised loss term of Eq. (A III-2). By doing so, the network tries to learn a meaningful feature representation without annotations. In the second stage, we apply the supervised loss of Eq. (A III-2) only on labeled images, enabling the network to fine-tune its representation and propagate acquired knowledge to the segmentation output space. For a fair comparison, we once again performed a grid search to select λ_1, λ_2 and λ_3 for this unsupervised setting, and report performance obtained with the best set of hyper-parameters.

The evolution of performance, in terms of average DSC on the ACDC validation set, is shown in Fig. 4.7. The mean and standard deviation are reported from three independent runs with different random seeds. We see that the Pretrain-Finetune strategy helps stabilize the training process and boosts segmentation performance by nearly 5.10% over Partial-Supervision. This confirms the ability of the proposed loss to learn useful representations when trained in an unsupervised setting. Nevertheless, our Joint-Optimization method outperforms this two-stage strategy by a significant margin, achieving the best validation score. This improvement is due to the fact that involving the labeled data into MI-based optimization helps the network find a more robust representation, thus improving the segmentation performance in an scarce-annotation setting.

4.7 Discussion and conclusion

We presented a novel semi-supervised method for medical images segmentation which regularizes a network by maximizing the MI between semantically-related feature embeddings, both globally and locally. The proposed global MI loss encourages the encoder to learn a transformation-invariant representation for unlabeled images. On the other hand, the local MI loss captures high-order dependencies between spatially-related embeddings, and preserves structure under perturbations of the input. By combining these two MI-based losses with a consistency term that promotes the alignment of cluster labels across different feature embeddings, the network can be effectively trained with limited supervision. We applied the proposed method to four challenging medical segmentation tasks with few annotated images. Experimental results showed our method to outperform recently-proposed semi-supervised approaches such as Mean Teacher and Entropy minimization, offering segmentation performance near to full supervision.

Standard loss functions for segmentation consider the prediction for different pixels as independent. An important advantage of our MI regularization losses is taking into consideration the structured nature of segmentation. Towards this goal, we maximize the MI on intermediate feature embeddings by using auxiliary projectors that map these continuous representations to a categorical distribution. While this provides an efficient way to estimate MI and promotes the grouping of semantically-related representations, other approximation techniques could also be explored. A possible alternative is adversarial contrastive learning (Bose, Ling & Cao, 2018), which employs an adversarially-learned sampler to find a reduced set of hard negative samples. Reducing the number of negative samples required to estimate MI could make approaches based on contrastive learning better-suited for the segmentation of large images and 3D scans. Another way to enhance the proposed method would be to incorporate priors on the distribution of segmentation labels. By maximizing MI, the proposed method indirectly favors balanced sizes for the segmented regions. When segmenting regions of very different sizes, better results could be achieved by constraining the marginal distribution of outputs (Hu *et al.*, 2017b). As future work, we could also validate the proposed method on multi-modal images, and large-scale segmentation benchmarks such as Cityscapes (Cordts *et al.*, 2016).

CHAPTER 5

SELF-PACED CONTRASTIVE LEARNING FOR SEMI-SUPERVISED MEDICAL IMAGE SEGMENTATION WITH META-LABELS

Jizong Peng[1] , Ping Wang[1] , Christian Desrosiers[1] , Marco Pedersoli[2]

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

Article accepted by *«Conference on Neural Information Processing Systems (NeurIPS) 2021»* in Dec. 2021.

5.1 Presentation

This chapter presents the article “Self-Paced Contrastive Learning for Semi-supervised Medical Image Segmentation with Meta-labels" by Peng, Pedersoli and Desrosiers, accepted by *the conference on Neural Information Processing Systems* for publication in Dec. 2021. The objective of this work is to develop a representation learning method based on self-pace mechanism, allowing to learn high quality representations. We show in this work various meta-labels, often available for free from volumetric images can be used to boost the semi-supervised segmentation not only in a pre-training stage, but also in a semi-supervised stage. Our method outperforms six state-of-the-art semi-supervised segmentation approaches and reaches a performance near full supervision given extremely small amount of labeled data.

5.2 Introduction

Since the emergence of deep learning (Krizhevsky *et al.*, 2012), there has been an active debate on the importance of pre-training neural networks. Precursor works (Erhan *et al.*, 2010) showed that pre-training a convolutional neural network with an unsupervised task (e.g., denoising autoencoders (Vincent, Larochelle, Bengio & Manzagol, 2008)) could lead to a better performance in the final supervised task. As the amount of labeled training data increased, thanks to large datasets like ImageNet (Deng *et al.*, 2009), it was however found that pre-training could actually hinder performance (Paine, Khorrami, Han & Huang, 2014). This makes sense in light of recent studies showing, for instance, that symmetries in large networks induce many equivalent local minima (Du, Lee, Li, Wang & Zhai, 2019; Nguyen & Hein, 2017; Soudry & Carmon, 2016) in which a pre-trained model can get stuck. Recently, contrastive learning has renewed the interest in unsupervised pre-training (Oord *et al.*, 2018). Several works (Chen *et al.*, 2020a; Chen, Kornblith, Swersky, Norouzi & Hinton, 2020b; Chen *et al.*, 2020c; He, Fan, Wu, Xie & Girshick, 2020; Zhao *et al.*, 2020c) have found that pre-training a model with a contrastive loss can improve its performance on a subsequent supervised training task, often outperforming a network with supervised pre-training on ImageNet. While this has reopened the debate on the benefit of pre-training, it offers little help for domains where data is scarce such as medical imaging. In medical imaging, not only are labels expensive since they come from highly-trained experts like radiologists, but images are also hard to obtain due to the need for costly equipment (e.g., MRI or CT scanner) and privacy regulations.

Over the last years, a breadth of semi-supervised learning approaches have been proposed for medical image segmentation, including methods based on attention (Min & Chen, 2018), adversarial learning (Zhang *et al.*, 2017c), temporal ensembling (Cui *et al.*, 2019; Yu *et al.*, 2019), co-training (Peng *et al.*, 2020a; Zhou *et al.*, 2019b), data augmentation (Chaitanya *et al.*, 2019; Zhao *et al.*, 2019a) and transformation consistency (Bortsova *et al.*, 2019). The common principle of these approaches is to add an unsupervised regularization loss using unlabeled images, which is optimized jointly with a standard supervised loss on a limited set of labeled images. Despite reducing significantly the amount of labeled data required for training, current semi-supervised learning methods still suffer from important drawbacks which impede their use in various applications. Thus, a large number of unlabeled images is often necessary to properly learn the regularization prior. As mentioned before, this may be impossible in medical imaging scenarios where data is hard to obtain. Moreover, these methods also need a sufficient amount of labeled data, otherwise the learning may collapse (Oliver *et al.*, 2018).

In a recent work, Chaitanya et al. (Chaitanya *et al.*, 2020) showed that unsupervised pre-training can be useful to learn a segmentation task with very few samples, by leveraging the meta information of medical images (e.g., the position of a 2D image in the 3D volume). While achieving impressive accuracy with as few as two volumes, this work has significant limitations. First, it relies on the strong assumption that the global or local representations of 2D images are similar if their locations within the volume or feature map are related. This assumption does not always hold in practice since volumes may not be well aligned, or due to the high variability of structures to segment. Second, it requires dividing the 2D images of a 3D volume in an arbitrary number of hard partitions that are contrasted, while the structure to segment typically varies gradually within the volume. Third, they do not exploit the full range of available meta data, for instance the patient ID or cycle phase of cardiac cine MRI, nor evaluate the benefit of combining several types of meta information in pre-training. Last, their approach leverages meta data only in pre-training, however this information could further boost performance if used while learning the final segmentation task, in a semi-supervised setting.

Our work addresses the limitations of current semi-supervised and self-supervised approaches for segmentation by proposing a novel self-paced contrastive learning method, which takes into account the noisiness of weak labels from meta data and exploits this data jointly with labeled images in a semi-supervised setting. The detailed contributions of this paper are as follows:

- We propose, to our knowledge, the first self-paced strategy for contrastive learning which dynamically adapts the importance of individual samples in the contrastive loss. This helps the model deal with noisy weak labels that arise, for instance, from misaligned images or splitting a 3D volume in arbitrary partitions.
- We demonstrate the usefulness of a contrastive loss on meta-data for improving the perfor-

mance of a final task, not only in pre-training but also as an additional loss in semi-supervised training.

- We show that combining multiple meta-labels in our self-paced contrastive learning framework can improve performance on the final task, compared to using them independently. Our results also demonstrate the benefit of combining contrastive learning with temporal ensembling to further boost performance.

We empirically validate our contributions on five well-known medical imaging datasets, and show the proposed approach to outperform the contrastive learning method of Chaitanya *et al.* (2020) as well as several state-of-the-art semi-supervised learning methods for segmentation (Peng, Pedersoli & Desrosiers, 2021a; Perone & Cohen-Adad, 2018; Vu *et al.*, 2019; Zhang, Cisse, Dauphin & Lopez-Paz, 2017b; Zhang *et al.*, 2017c). In the results, our approach obtains a performance close to fully supervised training with very few training scans.

5.3 Related work

We focus our presentation of previous works on two machine learning sub-fields that are most related to our current work: self-supervision, which includes contrastive learning, and self-paced learning.

Self-supervision and contrastive learning Self-supervision is a form of unsupervised learning where a pretext task is used to pre-train a model so to better perform a downstream task. Examples of pretext tasks are learning to sort a sequence (Lee, Huang, Singh & Yang, 2017; Xiong, Ren, Zeng & Urtasun, 2021), predicting rotations (Komodakis & Gidaris, 2018; Feng, Xu & Tao, 2019), solving a jigsaw puzzle (Misra & Maaten, 2020) and many others (Doersch & Zisserman, 2017; Dosovitskiy, Springenberg, Riedmiller & Brox, 2014; Kim *et al.*, 2020; Sermanet *et al.*, 2018; Zhang *et al.*, 2016). Most of these methods can improve performance on the downstream task when labeled data for training is scarce. However, when a large and general-enough dataset of labeled images like ImageNet (Deng *et al.*, 2009) is available, a simple supervised pre-training may sometimes provide better results (Huh, Agrawal & Efros, 2016; Morid, Borjali & Del Fiol, 2020). Recently, unsupervised contrastive learning (Chen *et al.*, 2020a; He *et al.*, 2020; Chen *et al.*, 2020b) was shown to boost performance even when learning a downstream task on a large dataset, and to improve over a model pre-trained in a supervised manner on a large dataset.

This approach is based on the simple idea of enforcing similarity in the representation of two examples from the same class (*positive* pairs), and increase representation dissimilarity on pairs from different classes (*negative* pairs) (Oord *et al.*, 2018; Tian *et al.*, 2019). In image analysis tasks, positive pairs are typically defined as two transformed versions of the same image, for instance using a geometric or color transformation, while negative ones are any other pair of images (Chen *et al.*, 2020a,b,c; He *et al.*, 2020).

In a recent work, Khosla *et al.* (Khosla *et al.*, 2020) showed that, when combined with true semantic labels, a contrastive learning based "pre-train and fine-tune" pipeline performed surprisingly well, outperforming conventional training with cross-entropy in some cases. The work in Zhao *et al.* (2020c) extended this idea to pixel-level tasks like semantic segmentation, clustering the representations of pixels in an image according to their labels. Applying a similar strategy to medical image segmentation, Chaitanya *et al.* (Chaitanya *et al.*, 2020) leveraged meta-labels from 3D scans in a local and global contrastive learning framework to improve performance when training with limited data. However, positive and negative pairs in their contrastive loss are defined using noisy "weak" labels, arising for example from misaligned images or an arbitrary partitioning of 3D volumes, which may lead to learning sub-optimal representations. The work in Zeng *et al.* (2021) mitigated this problem by imposing a maximum distance along the z axis between slices forming positive pairs. Moreover, existing approaches only exploit meta-labels in pre-training, instead of considering them jointly with labeled images in a semi-supervised strategy. Our work extends these approaches with a self-paced learning method that adapts the importance of positive pairs dynamically during training, and focuses the learning on the most reliable ones. In contrast to Chaitanya *et al.* (2020), we also exploit contrastive learning as regularization loss in semi-supervised training and show that further improvements can be achieved when combining it with a temporal ensembling strategy like Mean Teacher (Cui *et al.*, 2019; Yu *et al.*, 2019). A recent approach by Chen *et al.* (Chen *et al.*, 2020b) also performs contrastive learning for image recognition in a semi-supervised setting. In this approach, a large teacher model is trained with a unsupervised contrastive loss and then fine-tuned with a small fraction of labeled data. A student model is trained afterwards using a knowledge distillation technique. Hence, unlike our method, there is no joint optimization of the supervised and contrastive objectives. In this work, we show that significant improvements can be achieved by jointly optimizing these two objectives.

Self-paced learning A sub-category of curriculum learning (CL) (Bengio, Louradour, Collobert & Weston, 2009), self-paced learning (SPL) is inspired by the learning process of humans that gradually incorporates easy to hard samples in training (Kumar, Packer & Koller, 2010). The effectiveness of such strategy has been validated in various computer vision tasks (Wang *et al.*, 2018; Jiang *et al.*, 2014; Zhang, Meng, Zhao & Han, 2017a). Jiang *et al.* (Jiang *et al.*, 2014) proposed an SPL method considering both the difficulty and diversity of training examples, which outperformed conventional SPL methods that ignore diversity. Zhang *et al.* (Zhang *et al.*, 2017a) incorporated SPL in a DNN fine-tuning process for object detection, to cope with data ambiguity and guide the learning in complex scenarios. The usefulness of SPL when training with a limited budget or when the training data is corrupted by noise was also studied in recent work (Wu, Dyer & Neyshabur, 2021b). So far, the application of SPL to image segmentation remains limited. Wang *et al.* (Wang *et al.*, 2018) presented an SPL method for lung nodule segmentation, where the uncertainty of each sample prediction in the loss is controlled by the SPL regularizer. Similarly, a self-paced co-training method was proposed in Wang *et al.* (2021) for the semi-supervised segmentation of medical images. Related to our work, Liu *et al.* (2021c) proposed a margin preserving contrastive learning framework for domain adaptation that uses a self-paced strategy in self-training. Compared to this approach, which uses SPL *outside* the contrastive loss to select confident pseudo-labels for self-training, our method incorporates it *within* the loss via importance weights that are learned jointly with network parameters.

5.4 Proposed method

In this section, we present our self-paced contrastive learning approach for segmentation that leverages intrinsic meta information extracted from medical volumetric images. Our method effectively pre-trains a segmentation model with a self-paced variant of contrastive learning that is more robust to noisy annotations. The same approach is also used to further boost the segmentation accuracy in a semi-supervised setting, where only very limited pixel-wised annotations are used. In the following subsections, we detail the formulation of the proposed approach.

5.4.1 Contrastive learning with meta-labels

Given a batch of N images from a dataset $\mathcal{D}_u$ of unlabeled images, unsupervised contrastive learning approaches (Chen *et al.*, 2020a; Hjelm *et al.*, 2018; Oord *et al.*, 2018) aim at finding a feature extractor $f(\cdot)$ that gives similar representations for two augmented instances of the same image and different ones for those of two separate images, regardless of their true classes. This can be achieved by creating an augmented set of samples indexed by $i \in I \equiv 1, \ldots, 2N$, with two augmented samples for each original image in the batch. The following loss is then optimized:

$$\mathcal{L}_{\text{unsupCon}} = \frac{1}{2N} \sum_{i=1}^{2N} -\log \frac{\exp\left(\mathbf{z}_i^\top \mathbf{z}_{j(i)}/\tau\right)}{\sum_{a \in \mathcal{A}(i)} \exp\left(\mathbf{z}_i^\top \mathbf{z}_a/\tau\right)} \tag{5.1}$$

In this loss, $\mathbf{z}_i = \frac{f(\mathbf{x}_i)}{\|f(\mathbf{x}_i)\|}$[1] is the L_2-normalized representation of an image $\mathbf{x}_i$ in the augmented batch (i.e., the *anchor*) and $j(i)$ is the index of the other augmented sample from the same image (i.e., the *positive*). $\mathcal{A}(i) \equiv I \setminus \{i\}$ contains all indexes of the augmented set except i, and has a size of $2N-1$. Finally, τ is a small temperature factor that helps gradient descent optimization by smoothing the landscape of the loss.

This approach works well when a large dataset of unlabeled images is available. However, the number of available images is small in our case. To alleviate this problem, our contrastive learning framework also leverages meta-labels arising from the structure of the data. Following (Chaitanya *et al.*, 2020), we consider the 2D slices of a given set of M volumetric scans as our training data, and extract various meta-labels for each 2D image (e.g., patient ID, position of the slice in the volume, etc). More generally, we suppose that each image $\mathbf{x}_i$ has set of K meta-labels denoted as $y_i^k \in \{1,\ldots,C_k\}$, where C_k is the number of class labels for meta information $k \in \{1,\ldots,K\}$. The contrastive loss for the meta-label k is then defined as

$$\mathcal{L}_{\text{con}}^{k} = \frac{1}{2N}\sum_{i=1}^{2N}\frac{1}{|P^k(i)|}\sum_{j\in P^k(i)} \underbrace{-\log\frac{\exp\left(\mathbf{z}_i^{\top}\mathbf{z}_j/\tau\right)}{\sum_{a\in\mathcal{A}(i)}\exp\left(\mathbf{z}_i^{\top}\mathbf{z}_a/\tau\right)}}_{\ell_{ij}} \tag{5.2}$$

where $P^k(i) = \{j \in I \,|\, y_j^k = y_i^k\} \cup \{j(i)\}$ are the indexes of augmented samples with same label as $\mathbf{x}_i$, or coming from the same original image. By minimizing this loss, the feature extractor learns to group together representations with the same class and push away those from different ones.

In medical image segmentation, encoder-decoder based networks such as U-Net (Ronneberger *et al.*, 2015) and its variants are widely utilized thanks to their symmetric design and appealing performance on various dataset. Such network $F(\cdot)$ decomposes in two parts, an encoder $E(\cdot)$ that summarizes the global context of an input 2D slice into a low-dimensional representation,

[1] We omit the nonlinear projector head for the sake of simplification.

and a decoder $D(\cdot)$ that takes as input the representation and gradually recovers its spatial resolution using side information such as skip connections or pooling indexes. Previous work on contrastive learning showed that pre-training both the encoder and decoder separately helped the downstream segmentation task (Chaitanya *et al.*, 2020). In preliminary experiments, we found that pre-training the decoder gave marginal improvements and thus focused our method on the encoder. Specifically, we consider the features of the encoder as a single vector $E(\mathbf{x}) \in \mathbb{R}^d$ and use a shallow non-linear projector $g(\cdot)$ called *head* to obtain the final normalized embedding $\mathbf{z}_i = \frac{g(E(\mathbf{x}_i))}{\|g(E(\mathbf{x}_i))\|}$.

5.4.2 Self-paced learning to mitigate noisy meta-labels

The supervised contrastive loss of Equ. (5.2) can actually hurt the learning of representations in pre-training if the positive pairs are obtained with “weak” or noisy labels. For instance, if using patient ID as meta-label, we will force the encoder to cluster the representations of all 2D slices in a 3D volume, *including* those containing mainly background noise. Likewise, grouping together the slices in the same partition of two volumes hinders pre-training if the volumes are not fully aligned and/or their partitions cover different regions of the structure to segment.

To overcome this problem, we propose a self-paced strategy for contrastive learning which assigns an importance weight $w_{ij} \in [0, 1]$ to the specific loss of each positive pair (i, j), defined as l_{ij} in Equ. (5.2). A self-paced regularization term $R_\gamma(w_{ij})$, controlled by the learning pace parameter γ, is added to give a greater importance (i.e., larger w_{ij}) to pairs that are more confident (i.e., smaller ℓ_{ij}), and vice-versa. The learning pace γ is increased over training so that high-confidence pairs are considered in the beginning and then less-confident ones are gradually added as training progresses. We achieve this by defining the following self-paced contrastive loss optimized over both encoder parameters and importance weights:

$$\mathcal{L}^{k}_{\text{SP-con}} \;=\; \frac{1}{2N}\sum_{i=1}^{2N}\frac{1}{|P^k(i)|}\sum_{j\in P^k(i)} w_{ij}\,\ell_{ij} \;+\; R_\gamma(w_{ij}) \tag{5.3}$$

Following standard SPL approaches (Jiang *et al.*, 2014), we define the regularizer R_γ such that the weights are monotone decreasing with respect to the loss l_{ij} (i.e., harder examples are given less importance) and monotone increasing with respect to the learning pace (i.e., a larger γ increases the weights). In this work, we consider two regularizer functions, based on hard thresholding and linear imputation:

$$R^{\text{hard}}_\gamma(w_{ij}) \;=\; -\gamma\, w_{ij}\,; \qquad R^{\text{linear}}_\gamma(w_{ij}) \;=\; \gamma(\frac{1}{2}w^2_{ij} - w_{ij}). \tag{5.4}$$

Optimization process We minimize the loss in Equ. (5.3) by optimizing alternatively with respect to the encoder parameters Θ_E or importance weights w_{ij}, while keeping the other fixed. With fixed w_{ij}, we update Θ_E via stochastic gradient descent where the gradient is given by:

$$\nabla_{\Theta_E}\mathcal{L}^{k}_{\text{SP-con}} \;=\; \frac{1}{2N}\sum_{i=1}^{2N}\frac{1}{|P^k(i)|}\sum_{j\in P^k(i)} w_{ij}\,\nabla_{\Theta_E}\ell_{ij}. \tag{5.5}$$

As can be seen, the gradient of low-confidence pairs (i, j) will be scaled down by their weight w_{ij}, thus these pairs will contribute less to the learning. Then, given a fixed Θ_E we compute the optimal weights w^*_{ij} by solving the following problem:

$$w^*_{ij} \;=\; \operatorname*{arg\,min}_{w_{ij}\in[0,1]}\; w_{ij}\,\ell_{ij} \;+\; R_\gamma(w_{ij}) \tag{5.6}$$

The following proposition gives the optimal solution for the hard and linear SPL regularization strategies.

Proposition 1. *Given the definitions of R^{hard}_γ and R^{linear}_γ in Equ. (5.4), the closed-form solutions* *to Equ. (5.6) are given by*

$$w_{ij}^{\text{hard}} = \begin{cases} 1, & \textit{if } \ell_{ij} \leq \gamma \\ 0, & \textit{else} \end{cases} ; \qquad w_{ij}^{\text{linear}} = \max\left(1 - \frac{1}{\gamma}\ell_{ij},\, 0\right). \tag{5.7}$$

Proof. For the hard regularizer R_{γ}^{hard} the problem becomes

$$\min_{w_{ij}\in[0,1]} \; w_{ij}\,\ell_{ij} \; - \; \gamma\, w_{ij} \; = \; (\ell_{ij} - \gamma) w_{ij} \tag{5.8}$$

If $\ell_{ij} - \gamma \geq 0$, since we are minimizing, the optimum is obviously $w_{ij} = 0$. Else, if $\ell_{ij} - \gamma < 0$, the minimum is achieved for $w_{ij} = 1$. Combining these two results gives the hard threshold of Equ. (5.7).

A similar approach is used for the linear regularizer R_{γ}^{hard}. In this case, the problem to solve is

$$\min_{w_{ij}\in[0,1]} \; w_{ij}\,\ell_{ij} \; + \; \gamma(\frac{1}{2}w_{ij}^2 - w_{ij}) \; = \; \frac{\gamma}{2}w_{ij}^2 \; + \; (\ell_{ij} - \gamma)\, w_{ij} \tag{5.9}$$

If $l_{ij} \geq \gamma$, since $w_{ij} \geq 0$, the minimum is reached for $w_{ij} = 0$. Else, if $\ell_{ij} < \gamma$, we find the optimum by deriving the function w.r.t. w_{ij} and setting the result to zero, giving

$$w_{ij} \; = \; 1 \; - \; \frac{1}{\gamma}\ell_{ij}. \tag{5.10}$$

Since both γ and ℓ_{ij} are non-negative, we have that $w_{ij} \in [0, 1]$, hence it is a valid solution. Considering both cases simultaneously, we therefore get the linear rule of Equ. (5.7). □

These update rules in Equ. (5.7) can be explained intuitively. For a given γ, the hard threshold rule only considers confident pairs with $\ell_{ij} \leq \gamma$ and ignores the others. In contrast, the linear rule weighs each pair proportionally to γ and the inverse of ℓ_{ij}, emphasizing more confident ones.

Selecting the learning pace parameter One of the main challenges in self-paced learning methods is selecting the learning pace parameter γ. If γ is too small, all pairs will be ignored and there will be no learning. Conversely, if γ is too large, all pairs will be considered regardless of their confidence, which corresponds to having no self-paced learning. The following proposition provides insights on how to set this parameter during training.

Proposition 2. *The loss ℓ_{ij} related to a given pair (i, j) in the SPL objective of Equ. (5.3) is bounded by* $\log 2(N-1) - 2/\tau \ \leq \ell_{ij} \leq \ \log 2N + 2/\tau$*, where N is the batch size.*

Proof. We start by rewriting l_{ij} equivalently as

$$\ell_{ij} \ = \ \log \frac{\sum_{a\in\mathcal{A}(i)} \exp\left(\mathbf{z}_i^\top \mathbf{z}_a/\tau\right)}{\exp\left(\mathbf{z}_i^\top \mathbf{z}_j/\tau\right)} \ = \ \log\left(1 \ + \sum_{a\in\mathcal{A}(i)\setminus j} \frac{\exp\left(\mathbf{z}_i^\top \mathbf{z}_a/\tau\right)}{\exp\left(\mathbf{z}_i^\top \mathbf{z}_j/\tau\right)}\right) \tag{5.11}$$

Since the representation vectors $\mathbf{z}_i$ are L_2-normalized, their dot product is a cosine similarity falling in the range $[-1, 1]$. To minimize l_{ij}, we then need to minimize the dot product in the numerator inside the sum and maximize the one in the denominator. Using $|\mathcal{A}(i) \setminus j|= 2N-2$, we get

$$\ell_{ij}^{\min} \ = \ \log\left(1 \ + \ (2N-2)\,\frac{e^{-1/\tau}}{e^{1/\tau}}\right) \ = \ \log\left(1 \ + \ 2(N-1)e^{-2/\tau}\right) \ \geq \ \log 2(N-1) \ - \ 2/\tau. \tag{5.12}$$

Similarly, we maximize ℓ_{ij} by doing the opposite:

$$\ell_{ij}^{\max} \ = \ \log\left(1 \ + \ (2N-2)\,\frac{e^{1/\tau}}{e^{-1/\tau}}\right) \ = \ \log\left(1 \ + \ 2(N-1)e^{2/\tau}\right) \ \leq \ \log 2N \ + \ 2/\tau. \tag{5.13}$$

□

This proposition tell us that using $\gamma = \ell_{ij}^{\max}$ guarantees that all pairs are used in the loss, for both the hard and the linear SPL regularizers. Additionally, when using the hard regularizer R_γ^{hard}, $\gamma = \ell_{ij}^{\min}$ is the minimum learning pace so that at least one pair can be selected.

The complete SPL loss To exploit the information in all available meta-labels, our final loss combines the contrastive losses $\mathcal{L}\textsc{sp}\text{-con}^k$ for meta-labels $k = 1, \ldots, K$:

$$\mathcal{L}_{\textsc{sp}\text{-con}} = \sum_{k=1}^{K} \lambda_k \, \mathcal{L}^{k}_{\textsc{sp}\text{-con}} \tag{5.14}$$

Here, $\lambda_k \geq 0$ is a coefficient controlling the relative importance of the k^{th} meta-label in the final loss, which is determined by grid search on a separate validation set.

5.4.3 Semi-supervised segmentation with contrastive learning

In previous work (Chaitanya *et al.*, 2020), contrastive learning has mostly been used for pre-training the model. Here, we show that it can further boost results in a semi-supervised setting, where training is performed with a limited set of samples. In this setting, in addition to the unlabeled images $\mathcal{D}_u$, a small amount of pixelwise-annotated images $\mathcal{D}_l$ are also available. To incorporate the knowledge from meta information in a semi-supervised setting, we modify our self-paced contrastive loss as

$$\mathcal{L}_{\text{semi-sup}} = \mathcal{L}_{\text{sup}} + \lambda_{\text{reg}} \, \mathcal{L}_{\text{reg}} + \lambda_{\textsc{sp}} \, \mathcal{L}_{\textsc{sp}\text{-con}}, \tag{5.15}$$

where $\mathcal{L}_{\text{sup}}$ is the loss computed on labeled data (cross-entropy loss in our work), $\mathcal{L}_{\text{reg}}$ is the regularization loss normally used in semi-supervised approaches (in our experiments we use Mean Teacher) and $\mathcal{L}_{\text{SP-con}}$ is our self-paced contrastive loss. Last, λ_{reg} and λ_{sp} are weights balancing the different loss terms which are determined by grid search.

5.5 Experimental setup

To assess the performance of the proposed self-paced contrastive learning, we carry out extensive experiments on five benchmark datasets with different experimental settings. In this section, we briefly describe these datasets and give implementation details for our method. For further information, the reader can refer to the Supplementary Material.

5.5.1 Datasets

Five clinically-relevant benchmark datasets for medical image segmentation are used for our experiments: the Automated Cardiac Diagnosis Challenge (ACDC) dataset (Bernard *et al.*, 2018), the Prostate MR Image Segmentation 2012 Challenge (PROMISE12) dataset (Litjens *et al.*, 2014), and Multi-Modality Whole Heart Segmentation Challenge (MMWHS) dataset (Zhuang & Shen, 2016), as well as the Hippocampus and Spleen segmentation datasets from Antonelli *et al.* (2021). These datasets contain different anatomic structures and present different acquisition resolutions. For the contrastive loss, we exploit meta-labels on slice position and patient identity. Additionally, for ACDC, we consider the cardiac phase (i.e., systole or diastole) as a third source of meta-data. For all datasets, we split images into training, validation and test sets, which remain unchanged during all experiments. We train the model with only a few scans of the dataset as labeled data (the rest of the data is used without annotations as in a semi-supervised setting) and report results in terms of 3D DSC metric (Bertels *et al.*, 2019) on the test set. Details on the training set split, data pre-processing, augmentation methods and evaluation metrics can be found in the Supplementary Material.

For all datasets, we report the segmentation performance by varying the number of labeled scans across experiments. For the ACDC dataset, this number ranges from 1 to 4, representing 0.5% to 2% of all available data. For PROMISE12, we use 3 to 7 scans, representing 6% to 14% of the whole data. For MMWHS, we use 1 and 2 annotated scans, corresponding to 10% and

20% of the training data. We use 1 to 4 scans as annotated data for the Hippocampus dataset, representing 0.5% to 0.2% of the whole data, and 2 to 4 scans for the Spleen dataset, which corresponds to 5.7% to 11.4% of the whole available training data. Note that once randomly selected, those labeled volumes are fixed across the different experiments. Selecting labeled scans per experiment yielded significant variances (up to 11.25% in term of 3D DSC), as shown in the Supplementary Material. We include the segmentation results for both Hippocampus and Spleen datasets in the Supplementary Material.

5.5.2 Network architecture and optimization parameters

We use `PyTorch` (Paszke *et al.*, 2017) as our training framework and, following Chaitanya *et al.* (2020), employ the U-Net architecture (Ronneberger *et al.*, 2015) as our segmentation network. This 2D-based networks often works well for data with anisotropic acquisition resolutions. Moreover, it has a lower computational cost and require less GPU memory than its 3D counterparts. Network parameters are optimized using stochastic gradient descent (SGD) with a RAdam optimizer (Liu *et al.*, 2019). We provide the detailed training hyper-parameters in the Suppl. Material. For the pre-training process, we obtain representations by projecting the encoder's output to a vector of size 256, using a simple MLP network with one hidden layer and LeaklyReLU activation function, following Chen *et al.* (2020a). Our proposed self-paced contrastive learning objective, defined in Equ. (5.3), involves a learning pace parameter γ set as

$$\gamma = \gamma_{\text{start}} + (\gamma_{\text{end}} - \gamma_{\text{start}}) \times \left(\frac{\texttt{cur\_epoch}}{\texttt{max\_epoch}}\right)^p \tag{5.16}$$

where γ_{start}, γ_{end} are hyper-parameters controlling the importance weights in the beginning and the end of training, and `cur_epoch`, `max_epoch` are the current training epoch and the total number of training epochs respectively. p controls how fast γ increases during the optimization procedure.

5.6 Results

In this section, we first compare the hard and linear regularization strategy for SPL on ACDC. Then, we evaluate all components of our method in a comprehensive ablation study with a reduced set of training data on the different datasets. Finally, we compare our method with the most promising approaches for semantic segmentation in medical imaging, with reduced training data.

5.6.1 Hard vs. linear self-paced regularization

Table 5.1 reports the validation 3D DSC score for the hard and linear SPL, while training with different p in Equ. (5.16), and different numbers of annotated scans on the ACDC dataset. We observe that both SPL strategies (R_γ^{hard} and R_γ^{linear}) effectively help improve performance, however the linear strategy always leads to a higher improvement. This is because the hard strategy only employs binary weights, i.e., $w_{ij} \in \{0, 1\}$, whereas the linear strategy gradually increases w_{ij} and therefore provides a smoother optimization.

In Fig. 5.1 (a), we plot the value of γ over epochs for different values of p, and show in (b) the corresponding expectation of w_{ij} for all positive pairs. We observe that, for a large p, γ tends to be small for most of the training and mainly increases in the very end of the process, resulting in small w_{ij} for positive pairs. In contrast, when $p = 1/2$, we see a rapid increase of weights w_{ij} during training, which results in higher segmentation scores. This observation is inline with the findings from Platanios, Stretcu, Neubig, Poczos & Mitchell (2019) and Penha & Hauff (2019) on different tasks, where raising rapidly the self-paced learning rate in the first half of the training benefits the generalization performance. In the Suppl. Material, we also present a concrete evaluation of w_{ij} for three different scans during model optimization, corresponding to the four orange star markers in Fig. 5.1 (b). Since we found that R_γ^{linear} strategy works better than R_γ^{hard}, we will use R_γ^{linear} for all following experiments.

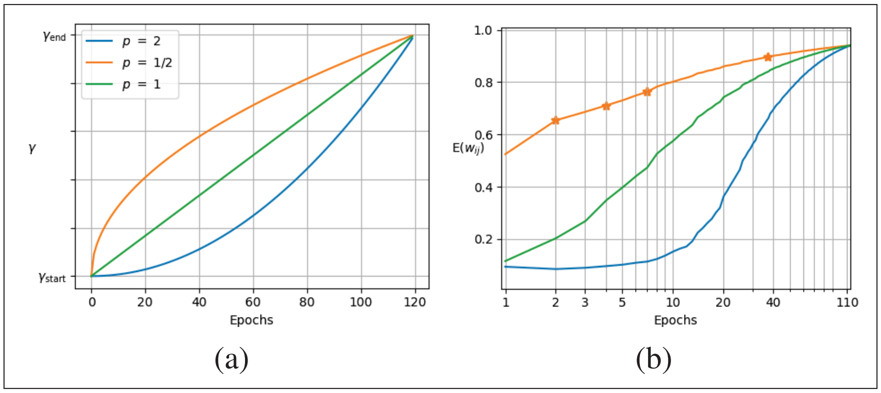


Figure 5.1 Self-paced strategy for γ. (a) Evolution of γ; (b) Expectation of w_{ij} over training epochs.

Table 5.1 3D DSC Performance on ACDC for hard and linear SP strategy and different values of p .

SP type	p	**ACDC**		
		1 scan	**2 scans**	**4 scans**
Baseline		57.53%	67.06%	75.64%
Linear	1/2	**74.40%**	**80.34%**	**81.86%**
	1	72.06%	79.54%	81.03%
	2	59.72%	70.36%	80.05%
Hard	1/2	64.42%	78.26%	80.07%
	1	**72.01%**	**79.80%**	**80.24%**
	2	71.86%	72.14%	76.19%

5.6.2 Ablation study

Table 5.2 summarizes the 3D DSC performance on test set for three datasets (ACDC, PROMISE12 and MMWHS) with very limited labeled data. At the top of the table, we report the number of labeled scans used and, for every result, also give in parenthesis the standard deviation computed with 3 different random seeds for parameter initialization. In the second and third columns of the table, we provide the loss used for the pre-training, if any, and the loss for the downstream training.

Upper and lower bounds We present results for a *Baseline* which uses only the annotated scans with cross-entropy as standard supervised loss $\mathcal{L}_{\text{sup}}$, and for *Full Supervision* where the same loss is used with all available data and associated annotations (175 for ACDC, 40 for Prostate and 10 for MMWHS). These two rows represent lower and upper bounds on the expected performance of the different variants of our approach.

Unsupervised contrastive loss We evaluate the performance of pre-training the network encoder with *Unsupervised Contrastive* loss as in Chen *et al.* (2020a), where two augmented versions of the same image are considered as a positive pair. In all datasets, this loss improves over our baseline model, although the improvement is limited because the amount of unlabeled data available is still reduced compared to the settings of previous work on unsupervised contrastive learning (Chen *et al.*, 2020a; He *et al.*, 2020; Chen *et al.*, 2020c,b; Zhao *et al.*, 2020c). We also add our self-paced learning strategy on top of this contrastive loss, and call this modified model *Unsupervised Contrastive + SP*. As meta-labels may be noisy, performance is increased in almost all experiments, especially when fewer labels are available.

Pre-training contrastive loss on meta-data We report the performance of a model pre-trained with a *Contrastive* loss on meta-labels. The meta-labels are 3D slice location $\mathcal{L}^1_{\text{con}}$, patient identity $\mathcal{L}^2_{\text{con}}$ and cardiac phase $\mathcal{L}^3_{\text{con}}$ (only for ACDC). We find that that slice position always gives the highest accuracy among all meta-labels, and largely outperforms the unsupervised contrastive loss. While all meta-labels increase performance compared to unsupervised contrastive loss, their combination leads to the best results in most cases.

Pre-training self-paced contrastive loss on meta-data Next, we evaluate the model pre-trained with a Self-Paced Contrastive loss on meta-labels (*SP-Con (pre-train)*). As with the unsupervised contrastive loss, the self-paced approach also successfully improves the segmentation quality compared to treating all positive pairs equally.

Semi-supervised We report the performance of a model without any pre-training, but using the unlabeled data during training with our proposed self-paced contrastive loss (*SP-Con (semi-sup)*). It can be seen that performance is inferior to Contrastive pre-training on meta-data, however the improvement is still quite relevant and, in most cases, superior to unsupervised pre-training.

Pre-trained and semi-supervised Our next subsection reports results for the combination of Self-Paced Contrastive learning used for both pre-training and semi-supervised training (*SP-Con (both)*). Although the loss is the same, its use during pre-training and as additional regularization in a semi-supervised setting brings additional improvements.

Pre-trained and semi-supervised with Mean Teacher We then evaluate the model using both pre-training and semi-supervised (as the previous setting) but with an additional Mean-Teacher for semi-supervision (*SP-Con (both)* + *Mean-Teacher*). By combining our approach with a simple Mean-Teacher method, our results on all datasets are further boosted, approaching the performance of fully supervised training but using a very low number of annotated scans. This is the model that is used in the comparison with the state-of-the-art.

5.6.3 Comparison with the state-of-the-art

We compare our method with other approaches that aim to improve training with few annotated images/scans. Table 5.3 presents results in terms of 3D DSC score for approaches based on data augmentation (Zhang *et al.*, 2017b), pre-training the weights on both encoder and decoder of the model (Chaitanya *et al.*, 2020) and various semi-supervised learning methods (Vu *et al.*, 2019; Zhang *et al.*, 2017c; Perone & Cohen-Adad, 2018; Peng *et al.*, 2021a). As with the ablation study, we report results for the ACDC, PROMISE12 and MMWHS datasets. A detailed explanation of the experimental setup of each method and results for the other two datasets can be found in the Supplementary Material.

To have a fair comparison, for all methods, we used grid search on the validation set to tune the

Table 5.2 3D DSC performance (and standard deviation) for different components and approaches on three medical image datasets with a few labelled scans.

Method	Pretrain	Train	ACDC			PROMISE12			MMWHS	
			1 scan	**2 scans**	**4 scans**	**3 scans**	**5 scans**	**7 scans**	**1 scan**	**2 scans**
Baseline	–	$\mathcal{L}_{\text{sup}}$	57.53 (1.18)	67.06 (0.68)	75.64 (0.15)	35.02 (2.59)	59.03 (1.25)	72.81 (1.08)	70.94 (0.84)	80.25 (0.38)
Full Supervision (*all labels*)	–	$\mathcal{L}_{\text{sup}}$	88.06 (0.20)			89.70 (0.51)			88.27 (1.23)	
Unsup. Con.	$\mathcal{L}_{\text{con}}^{\text{unsup.}}$	$\mathcal{L}_{\text{sup}}$	65.14 (2.53)	72.88 (3.00)	76.56 (1.34)	38.74 (8.47)	60.99 (5.20)	75.28 (1.72)	74.12 (1.45)	80.57 (2.50)
Unsup. Con. + SP	$\mathcal{L}_{\text{SP}}^{\text{unsup.}}$	$\mathcal{L}_{\text{sup}}$	67.35 (1.98)	75.11 (0.92)	76.87 (0.84)	40.21 (6.63)	67.37 (0.99)	75.14 (0.50)	74.30 (1.81)	80.71 (0.83)
Contrastive	$\mathcal{L}_{\text{con}}^{1}$	$\mathcal{L}_{\text{sup}}$	70.57 (0.96)	78.59 (0.79)	79.60 (0.49)	57.44 (4.89)	75.21 (1.94)	80.02 (1.28)	76.53 (1.79)	83.05 (2.68)
	$\mathcal{L}_{\text{con}}^{2}$		63.63 (1.80)	73.30 (1.25)	76.83 (0.91)	55.50 (3.83)	69.95 (1.06)	78.93 (0.63)	74.71 (0.29)	82.41 (0.33)
	$\mathcal{L}_{\text{con}}^{3}$		64.52 (1.34)	76.81 (0.97)	77.66 (0.56)	–	–	–	–	–
SP-Con (*pre-train*)	$\mathcal{L}_{\text{SP}}^{1}$	$\mathcal{L}_{\text{sup}}$	73.99 (1.27)	81.01 (1.44)	82.83 (0.26)	58.81 (2.35)	75.28 (1.49)	80.71 (1.27)	77.20 (0.87)	82.87 (0.39)
	$\mathcal{L}_{\text{SP}}^{2}$		69.26 (1.69)	76.34 (0.60)	78.34 (0.42)	56.80 (1.59)	69.75 (0.47)	79.02 (0.18)	76.67 (0.48)	83.10 (1.51)
	$\mathcal{L}_{\text{SP}}^{3}$		65.18 (1.50)	79.05 (.026)	81.04 (0.16)	–	–	–	–	–
SP-Con (*semi-sup*)	–	$\mathcal{L}_{\text{SP}}^{1}$	67.34 (0.74)	73.74 (0.51)	77.27 (0.12)	54.50 (1.53)	70.49 (1.33)	76.95 (0.81)	73.82 (0.68)	81.63 (0.39)
		$\mathcal{L}_{\text{sup}}+\ \mathcal{L}_{\text{SP}}^{2}$	60.82 (0.98)	68.06 (1.09)	77.10 (0.35)	41.67 (1.59)	61.04 (1.46)	75.98 (0.98)	73.43 (1.33)	78.08 (1.88)
		$\mathcal{L}_{\text{SP}}^{3}$	62.52 (0.46)	68.39 (0.26)	77.24 (0.17)	–	–	–	–	–
SP-Con (*both*)	$\mathcal{L}_{\text{SP}}^{1}$	$\mathcal{L}_{\text{SP}}^{1}$	75.66 (1.94)	80.37 (0.36)	82.35 (0.58)	68.79 (2.63)	77.38 (1.90)	80.55 (0.75)	76.58 (1.00)	82.69 (0.39)
	$\mathcal{L}_{\text{SP}}^{2}$	$\mathcal{L}_{\text{sup}}+\ \mathcal{L}_{\text{SP}}^{2}$	70.47 (0.93)	76.58 (0.45)	78.37 (0.22)	56.68 (2.64)	72.21 (1.32)	77.28 (1.86)	75.33 (0.62)	82.39 (0.27)
	$\mathcal{L}_{\text{SP}}^{3}$	$\mathcal{L}_{\text{SP}}^{3}$	70.08 (0.96)	78.70 (0.51)	80.19 (0.28)	–	–	–	–	–
SP-Con (*both*) + Mean Teacher	$\mathcal{L}_{\text{SP}}^{1}$	$\mathcal{L}_{\text{SP}}^{1}$	78.76 (0.26)	82.14 (0.19)	84.42 (0.18)	74.06 (1.13)	82.50 (0.91)	84.14 (0.35)	78.82 (0.34)	84.90 (0.58)
	$\mathcal{L}_{\text{SP}}^{2}$	$\mathcal{L}_{\text{sup}}\ \ \mathcal{L}_{\text{SP}}^{2}$	75.30 (0.68)	79.67 (0.26)	82.65 (0.32)	61.39 (1.33)	77.74 (1.07)	83.92 (0.39)	75.94 (0.90)	84.57 (0.61)
	$\mathcal{L}_{\text{SP}}^{3}$	$\mathcal{L}_{\text{MT}}\ +\ \mathcal{L}_{\text{SP}}^{3}$	73.94 (0.54)	81.29 (0.09)	83.21 (0.05)	–	–	–	–	–
	$\mathcal{L}_{\text{SP}}^{1-3}$	$\mathcal{L}_{\text{SP}}^{1-3}$	**79.80** (0.33)	**83.20** (0.25)	**84.84** (0.15)	**74.47** (0.36)	**83.78** (0.30)	**84.52** (0.17)	**78.97** (0.52)	**84.87** (0.11)

Table 5.3 3D DSC performance (and standard deviation) of our method and other approaches on three medical image datasets with few labelled scans. Bold red-colored values are the best performing methods, underlined blue-colored ones correspond to the second best performing method.

Method	ACDC			PROMISE12			MMWHS	
	1 scan	**2 scans**	**4 scans**	**3 scans**	**5 scans**	**7 scans**	**1 scan**	**2 scans**
Entropy Min. (Vu *et al.*, 2019)	60.47 (1.03)	69.81 (0.99)	76.19 (1.21)	53.47 (5.70)	65.66 (0.42)	73.52 (2.71)	72.28 (0.58)	78.39 (1.54)
Mix-up (Zhang *et al.*, 2017b)	60.87 (1.28)	67.45 (1.04)	76.18 (0.49)	41.38 (2.80)	64.55 (1.93)	73.56 (0.61)	71.50 (0.54)	80.12 (0.84)
Adv. Training (Zhang *et al.*, 2017c)	63.05 (0.80)	70.68 (0.27)	75.89 (0.94)	61.58 (2.10)	71.00 (1.20)	81.05 (1.34)	73.47 (1.42)	80.40 (0.93)
Mean Teacher (Perone & Cohen-Adad, 2018)	62.85 (0.67)	72.84 (0.22)	79.12 (0.08)	52.96 (1.97)	68.38 (2.04)	77.37 (0.87)	72.36 (1.35)	81.01 (0.57)
Discrete MI (Peng *et al.*, 2021a)	69.27 (1.41)	77.74 (0.42)	80.06 (0.24)	47.77 (3.58)	68.29 (2.35)	77.63 (1.13)	72.38 (1.04)	82.45 (1.36)
Contrastive (Chaitanya *et al.*, 2020)	70.05 (2.66)	79.11 (2.02)	81.25 (2.15)	61.15 (2.95)	74.62 (1.69)	80.08 (1.39)	76.45 (0.62)	82.93 (0.42)
Our Method	**79.80** (0.33)	**83.20** (0.25)	**84.84** (0.15)	**74.47** (0.36)	**83.78** (0.30)	**84.52** (0.17)	**78.97** (0.52)	**84.87** (0.11)

hyper-parameters. For most methods, the improvement with respect to the baseline trained with only the supervised loss is quite limited and varies depending on the dataset and the number of annotated scans used. For instance, *Adversarial training* performs quite well on the PROMISE12 dataset (scans in this dataset exhibit more variability in terms of intensity contrast), but not so well on ACDC. Likewise, *Mean-Teacher* does not perform well for 1 or 2 annotated scans in ACDC but, when increasing the scans to 4, it outperforms most of the other methods. The global and local *Contrastive* loss using meta-data manages to obtain an excellent improvement on all datasets. However, our approach still yields substantial improvements with respect to that method. This is due to our proposed self-paced learning strategy, as well as the combined use of the contrastive loss for pre-training and semi-supervised learning.

5.7 Discussion and conclusion

In this paper, we proposed a technique based on contrastive loss with meta-labels that can highly improve the performance of a medical image segmentation model when training data is scarce. It was shown that, with a reduced amount of unlabeled images, unsupervised contrastive loss is not very effective. Instead, in the context of medical images, additional meta-data is freely available and, if properly used, can greatly boost performance. We presented results on five well-known medical image datasets and have shown that the accuracy of the contrastive loss with meta-labels can be boosted by the use of self-paced learning. Our self-paced contrastive learning method can be used during pre-training as well as a regularization loss during semi-supervised training, and the combination of the two can further boosts results. Finally, we have compared our approach with the state-of-the-art in semi-supervised learning, and have shown that the simple combination of our approach with multiple meta-data and a simple semi-supervised approach as Mean Teacher is more effective than previous approaches. While using a few scans, our method can approach fully supervised training.

5.8 Social impact and limitations

The proposed method can have an effective and practical impact in terms of medical imaging analysis in hospitals and health centers. As shown in our experiments, it produces an accurate medical image segmentation with a very reduced set of annotated data. This has the potential of helping radiologists and other clinicians using medical images, which can in turn contribute to a better diagnosis and reduced costs. While our empirical evaluation has shown excellent results with very limited data, using fewer annotated images also increases chances of over-fitting potential outliers in the data that may lead to erroneous or misleading results. A further study on the reliability of medical image segmentation with reduced images is therefore recommended.

The proposed method also has technical limitations that could be addressed in future work. First, although it employs three meta labels to pretrain a 2-D network encoder, these meta labels cannot be easily transferred to pretrain a 3-D segmentation architecture. Moreover, our method also requires the tuning of several hyperparameter, which can be computationally expensive. Last, while our approach pretrains the encoder, pre-training the dense features from decoder layers could also be considered. This is however challenging since no meta-labels at the pixel level are available.

CHAPTER 6

BOUNDARY-AWARE INFORMATION MAXIMIZATION FOR SELF-SUPERVISED MEDICAL IMAGE SEGMENTATION

Jizong Peng[1] , Ping Wang[1] , Marco Pedersoli[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] Department of Automated Production, École de Technologie Supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3

Article submitted to *«International Conference on Machine Learning (ICML)»* in Jan. 2022.

6.1 Presentation

This chapter presents the article “Boundary-aware Information Maximization for Self-supervised Medical Image Segmentation" by Peng, Wang, Pedersoli and Desrosiers, submitted to *International Conference on Machine Learning (ICML)* in Jan. 2022. In this work, we propose a novel unsupervised pre-training framework that avoids the drawback of contrastive learning. Our framework consists of two principles: unsupervised over-segmentation as a pre-train task using mutual information maximization and boundary-aware preserving learning. Experimental results on two benchmark medical segmentation datasets reveal our method’s effectiveness in improving segmentation performance when few annotated images are available.

6.2 Introduction

Supervised deep learning approaches have achieved outstanding performance in a wide range of segmentation tasks (Ronneberger *et al.*, 2015; Badrinarayanan *et al.*, 2017; Chen *et al.*, 2017). However, these approaches often require a large amount of labeled images which are difficult to obtain for medical imaging applications (Cheplygina, de Bruijne & Pluim, 2019; Peng & Wang, 2021). Unsupervised representation learning (Jing & Tian, 2020; Liu *et al.*, 2021b) has emerged as an effective technique to boost the performance of a segmentation model without the need for annotated data. In such technique, a model is pre-trained to perform a given pretext task, for example puzzle-solving (Noroozi & Favaro, 2016; Taleb *et al.*, 2021), rotation prediction (Komodakis & Gidaris, 2018), colorization (Zhang *et al.*, 2016) or contrastive-based instance discrimination (Hjelm *et al.*, 2018; Chen *et al.*, 2020a; He *et al.*, 2020), and then fine-tuned with a small set of labeled examples. Among these self-supervised methods, contrastive learning has become a prevailing strategy for pre-training medical image segmentation models (Chaitanya *et al.*, 2020; Zeng *et al.*, 2021; Peng, Wang, Desrosiers & Pedersoli, 2021b). The core idea of this strategy is to learn, without pixel-wise annotations, an image representation which can discriminate related images (e.g., two transformations of the same image) from non-related ones. Most contrastive learning approaches for segmentation apply a contrastive loss on the *global* representation of images, which typically corresponds to the features produced by the network's encoder. Experimental results have shown that pre-training the encoder with this loss and then fine-tuning the whole network with few labeled examples can lead to significant improvements (Peng *et al.*, 2021b).

Recent works have also demonstrated the benefit of using contrastive learning on the decoder's feature maps during pre-training (Chaitanya *et al.*, 2020; Peng *et al.*, 2021b). In this case, the contrastive loss is applied at each position of the feature map, which helps learn a *local* representation of the image. However, choosing the pairs of positive and negative examples that need to be contrasted is more challenging for these dense feature maps without pixel-wise annotations. Firstly, the meta-information in medical data (e.g., subject ID, slice position, etc.) is typically found at the image level, and is therefore not applicable to local contrastive learning. To tackle this problem, current methods usually adopt a stride sampling strategy where, for a given anchor position in the feature map, local representations located at a sufficient distance are regarded as negative, while those that are close but obtained under different image transforms are considered as positive (Chaitanya *et al.*, 2020). As we show in our experiments (see Section

6.5.1), this weak spatial prior unfortunately leads to low improvements when used in pre-training. Another problem comes from the fact that medical images for segmentation are often dominated by non-informative background regions, which reduces the effectiveness of local contrastive learning in this setting. Additionally, standard contrastive learning techniques such as (Hjelm *et al.*, 2018) typically need large batch sizes to have a sufficient amount of high-quality negative example pairs. This constraint can be hard to meet in the case of learning dense features. Despite important efforts, the improvement brought by *local* contrastive learning in medical image segmentation remains relatively marginal (Chaitanya *et al.*, 2020).

In this paper, we propose a boundary-aware information maximization approach for unsupervised representation learning and experimentally demonstrate its usefulness for medical image segmentation. Our approach focuses on the dense features in the decoder of a segmentation network, and seeks to group them into clusters that correspond to meaningful regions in the image. The proposed learning objective is based on the Information Invariant Clustering (IIC) method (Ji *et al.*, 2019), but overcomes three major drawbacks of this method: i) its optimization difficulty, caused in part by minimizing the entropy of cluster assignments, which often leads to sub-optimal solutions; ii) its lack of clustering consistency for different random transformations; iii) the poor correspondence of clusters obtained by this method with region boundaries in the image. As illustrated in Fig. 6.1, our boundary-aware information maximization approach learns clusters that better correspond to relevant anatomical structures of the image. This is achieved by improving IIC in two important ways. First, we augment the learning objective of IIC, which maximizes the mutual information of local feature embeddings for two different transformations of the same image, to make the joint cluster probability close to a uniform diagonal matrix. This improves optimization and leads to clusters that are well balanced and also consistent across different image transformations. Second, we propose a boundary-aware loss based on the cross-correlation between the spatial entropy of clusters and image edges, which helps the learned cluster be more representative of important regions in the image. Our experimental results reveal this loss to be especially effective for the segmentation of regions with irregular shape.

Compared to contrastive learning, our method does not require to compute positive or negative pairs, and does not need a sophisticated sampling mechanism or large batch sizes. Through an extensive set of experiments involving four different medical image segmentation tasks, we demonstrate the high effectiveness of our unsupervised representation learning method for pre-training a segmentation model, before fine-tuning it with few labeled images. Our results show the proposed method to outperform by a large margin several state-of-the-art self-supervised and semi-supervised approaches for segmentation, and to reach a performance close to full supervision with only a few labeled examples.

6.3 The proposed method

In unsupervised representation pre-training, we are given a set of N images $\mathcal{D} = \{x_i\}_{i=1}^{N}$, with $x_i \in \mathbb{R}^{\Omega}$, where Ω is the image space. We seek to learn a useful representation by pre-training a deep segmentation network $f_{\theta}(\cdot) = \phi_{\mathrm{dec}}(\phi_{\mathrm{enc}}(\cdot))$ comprised of encoder $\phi_{\mathrm{enc}}(\cdot)$ and a decoder $\phi_{\mathrm{dec}}(\cdot)$. In our setting, a good representation can boost segmentation performance when fine-tuning the whole network with very limited labeled data. To help understand our method, we summarize in Table 6.1 the main notations used in the paper.

Our representation operates on dense embeddings taken from some intermediate layer of the decoder. Our goal is to group these local embeddings into clusters reflecting meaningful anatomical structures in the input images, without requiring any labels. Three separate loss functions are used to achieve this goal. The first loss maximizes the MI between corresponding local feature embeddings obtained from an input image and its transformed version. Since computing MI between continuous variables is complex, as in recent works (Ji *et al.*, 2019; Peng *et al.*, 2021a), we project features to a discrete space representing clusters, where MI is

Table 6.1 Notations used in the paper

Image dataset:	$\mathcal{D} = \{x_i \in \mathbb{R}^{\Omega}\}_{i=1}^{N}$
Pixel index:	$\Omega = [1, \ldots, W \times H]$
Dense embedding index:	$\Omega' = [1, \ldots, W/N \times H/N]$
$(K-1)$-simplex:	$\Delta^K = \{\mathbf{p} \in [0,1]^K, \sum_k p_k = 1\}$
Dense embedding:	$\mathbf{s} = s(x) \in \mathbb{R}^{\Omega' \times C}$
Cluster projection:	$g(\mathbf{s}) \in \Delta^{\Omega' \times K}$
Image transform:	$\mathcal{T}(\cdot)$
Cluster probabilities:	$\widehat{\mathbf{p}}_i = g(s(\mathcal{T}(x_i))),\ \widetilde{\mathbf{p}}_i = g(\mathcal{T}(s(x_i)))$
Cluster marginals:	$\widehat{\mathbf{p}} = \frac{1}{N}\sum_{i=1}^{N}\widehat{\mathbf{p}}_i,\ \widetilde{\mathbf{p}} = \frac{1}{N}\sum_{i=1}^{N}\widetilde{\mathbf{p}}_i$
Joint distribution:	$\mathcal{P}_{\text{joint}} = \frac{1}{N}\sum_{i=1}^{N}\widehat{\mathbf{p}}_i \cdot \widetilde{\mathbf{p}}_i^{\intercal}$
Entropy:	$\mathcal{H}(X) = -\mathbb{E}_X[\log p(X)]$
Joint entropy:	$\mathcal{H}_{\text{joint}}(X,Y) = -\mathbb{E}_{X,Y}[\log p(X,Y)]$

easy to obtain. However, maximizing the MI between cluster assignments has two important drawbacks. Firstly, it assumes that the number of clusters is known in advance and that these clusters are balanced (i.e., represent regions of the same size in the image). Secondly, as it involves minimizing entropy, the direct optimization of MI often leads to poor local minima, as the network becomes quickly confident in cluster assignments that are not useful (see the clusters in Figure 6.1, for α=0). Our first loss term addresses this problem by combining two complementary objectives: 1) minimizing the entropy of the cluster assignment joint distribution, which encourages clusters to be balanced and confident but is also flexible to ignore some irrelevant clusters; 2) making the matrix of this joint distribution close to a diagonal matrix, which helps the optimization avoid poor minina and learn a better representation. Another problem with the simple MI maximization approach for unsupervised representation learning is that the clusters may not align with geometric cues such as edges in the input image. In the second loss of our model, we tackle this problem by forcing the regions with high cluster entropy, which corresponds to boundaries between clusters, to be correlated with edges in the image. Finally, to help the network capture the global context of images, we include a contrastive learning loss that exploits available meta-labels (e.g., slice position in a MRI volume) to make

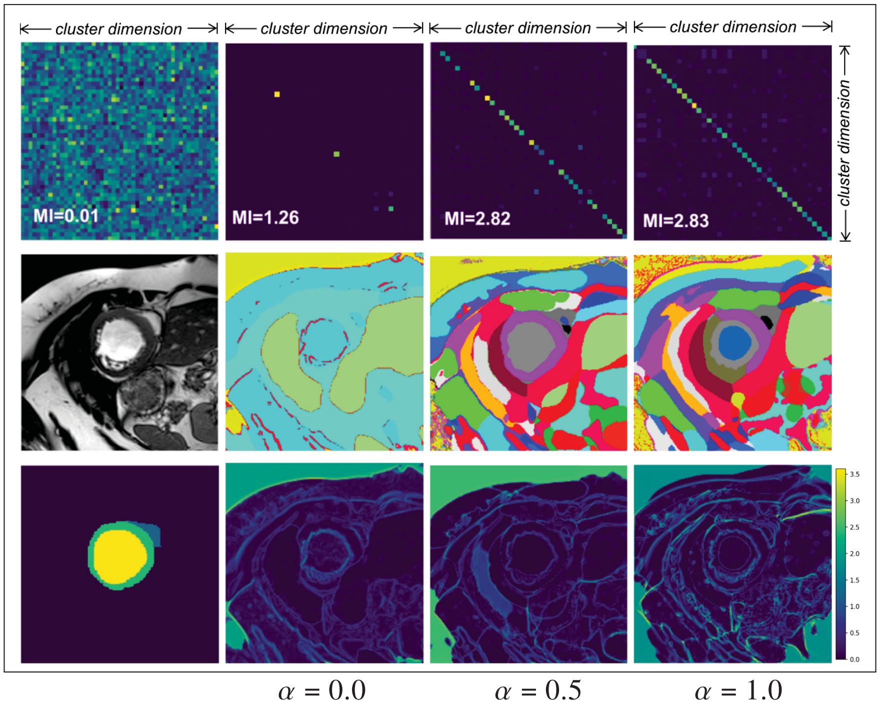


Figure 6.1 Influence of α on joint matrix $\mathcal{P}_{\text{joint}}$ (first row), cluster assignment (second row), and the uncertainty of the cluster (third row). The first column shows the joint matrix before optimization, the input image and the groud-truth segmentation respectively. Using a combination of MI and cross-entropy loss ($\alpha = 0.5$) provides the most meaningful unsupervised segmentation.

the global features obtained by the encoder f_E similar for images with the same meta-label. We present a conceptual diagram of our proposed method in Fig. IV-1 of the Appendix and detail the three loss functions in the following sub-sections.

6.3.1 Improved MI-based loss for *dense* pre-training

We seek to cluster the dense embeddings in feature maps **s** taken from a given hidden layer of the decoder $\phi_{\text{dec}}(\cdot)$. Following Peng *et al.* (2021a), we use mutual information maximization to perform clustering. The MI between two random variables X and Y (i.e., the cluster assignment for two images) corresponds to the KL divergence between their joint distribution $p(X, Y)$ and

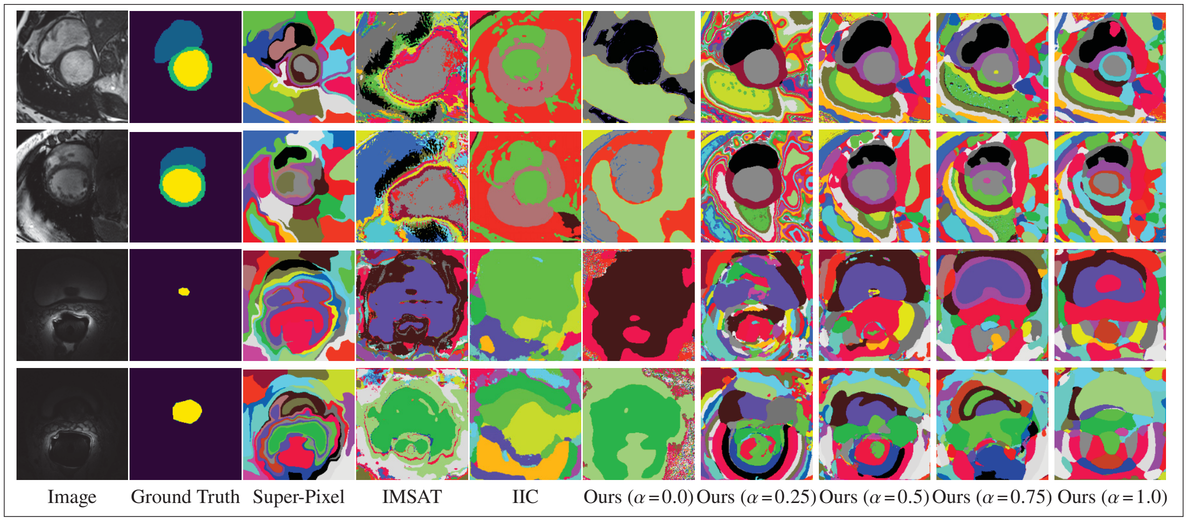


Figure 6.2 Visual inspection of cluster assignments for unsupervised pre-training

the product of their marginal distributions $p(X)$ and $p(Y)$:

$$I(X,Y) \;=\; (p(X,Y)\,||\,p(X)\,p(Y)) \tag{6.1}$$

Alternatively, MI can also be defined as the difference between the combined entropy of marginals and the entropy of the joint distribution:

$$\begin{aligned} I(X,Y) \;&=\; \mathcal{H}(X)+\mathcal{H}(Y)-\mathcal{H}(X,Y) \\ &=\; -\mathbb{E}_X[\log \mathbb{E}_Y[p(X,Y)]] - \mathbb{E}_Y[\log \mathbb{E}_X[p(X,Y)]] \\ &\qquad + \mathbb{E}_{X,Y}[\log p(Y,X)] \end{aligned} \tag{6.2}$$

where $\mathcal{H}(\cdot)$ is the entropy of the variable. This definition reveals that maximizing MI leads to high-entropy (uniform) distributions for X and Y, thus avoiding trivial solutions assigning all examples to a single cluster. It also results in a low entropy of the joint distribution, corresponding to confident cluster assignments.

Let $\widehat{\mathbf{p}}_i = g(s(\mathcal{T}(x_i)))$ and $\widetilde{\mathbf{p}}_i = g(\mathcal{T}(s(x_i)))$ be cluster probabilities in feature maps from a given layer of the decoder, obtained by applying a random transformation $\mathcal{T}(\cdot)$ on the input image x_i or the feature maps $s(x_i)$. Function g is a $1{\times}1$ convolutional layer followed by a K-way softmax projecting the feature maps to a distribution over K clusters. As in IIC, we estimate the joint distribution using the average outer product between cluster probabilities $\widehat{\mathbf{p}}_i$ and $\widetilde{\mathbf{p}}_i$:

$$\mathcal{P}_{\text{joint}} = \frac{1}{N} \sum_{i=1}^{N} \widehat{\mathbf{p}}_i \cdot \widetilde{\mathbf{p}}_i^{\mathsf{T}}. \tag{6.3}$$

$\mathcal{P}_{\text{joint}}$ thus has a dimensionality of $K{\times}K$, and $\mathcal{P}_{\text{joint}}^{(j,k)}$ is the joint probability of assigning $s(\mathcal{T}(x_i))$ to cluster j and $\mathcal{T}(s(x_i))$ to cluster k. Following Equ. (6.2), the MI between the corresponding random variables X and Y can be written as

$$I(X,Y) \;=\; \mathcal{H}(\widehat{\mathbf{p}}) + \mathcal{H}(\widetilde{\mathbf{p}}) - \mathcal{H}(\mathcal{P}_{\text{joint}}) \tag{6.4}$$

where $\widehat{\mathbf{p}} = \frac{1}{N} \sum_i \widehat{\mathbf{p}}_i$, $\widetilde{\mathbf{p}} = \frac{1}{N} \sum_i \widetilde{\mathbf{p}}_i$ are the cluster marginals which can computing by summing over the rows or columns of the joint distribution matrix. Maximizing the entropy of marginals encourages the network to assign an even number of samples to each cluster, and avoids trivial solutions where most clusters are empty. On the other hand, minimizing the entropy of the joint, $\mathcal{H}(\mathcal{P}_{\text{joint}})$, forces the network to have confident cluster assignments.

Clustering the dense embeddings by maximizing MI poses two optimization problems. First, since we are minimizing the entropy of the joint, the network can get stuck in confident but incorrect cluster assignments which remain the same throughout optimization. Another problem stems from the fact that MI is invariant to the ordering of clusters, hence any permutation of the joint distribution matrix yields an equivalent solution. The challenge of maximizing MI is illustrated in the first row of Fig. 6.1, where the left-most image is the initial joint matrix $\mathcal{P}_{\text{joint}}$ before optimization and the one in the second column ($\alpha = 0$) is $\mathcal{P}_{\text{joint}}$ after maximizing MI. We see that only a few clusters are actually used, making the entropy of marginals low and therefore also the MI.

To alleviate these problems, we consider the entropy of the joint distribution, given by

$$\mathcal{H}(\mathcal{P}_{\text{joint}}) \;=\; -\sum_{j=1}^{K}\sum_{k=1}^{K}\mathcal{P}_{\text{joint}}^{(j,k)}\log\mathcal{P}_{\text{joint}}^{(j,k)}. \tag{6.5}$$

A solution where cluster assignments are balanced, confident and perfectly consistent across transformations, would give a joint distribution matrix with diagonal elements equal to $1/K$ and off-diagonal elements to 0. To guide the optimization toward this desirable solution, we introduce a pseudo-label of the joint matrix $\mathcal{P}_{\text{pseud}} = \frac{1}{K}\mathrm{I}_K$, where I_K is the $K{\times}K$ identity matrix, and modify the entropy of the joint as follows:

$$\mathcal{H}'_{\alpha}(\mathcal{P}_{\text{joint}}) = -\sum_{j=1}^{K}\sum_{k=1}^{K}((1-\alpha)\cdot\mathcal{P}_{\text{joint}}^{(j,k)} + \alpha\mathcal{P}_{\text{pseud}}^{(j,k)})\log\mathcal{P}_{\text{joint}}^{(j,k)} \tag{6.6}$$

In this modified formulation, α is a mixing coefficient ranging from 0 to 1. If α equals to 0, $\mathcal{H}'_{\alpha}(\mathcal{P}_{\text{joint}})$ reduces to $\mathcal{H}(\mathcal{P}_{\text{joint}})$, while $\alpha = 1$ corresponds to a cross-entropy loss guiding the joint matrix towards the pre-defined diagonal solution $\mathcal{P}_{\text{pseud}}$.

Since the joint distribution matrix is computed over a batch of examples, minimizing the cross-entropy between $\mathcal{P}_{\text{pseud}}$ and $\mathcal{P}_{\text{joint}}$ is not the same as minimizing the cross-entropy between individual cluster assignments $\widehat{\mathbf{p}}_i$ and $\widetilde{\mathbf{p}}_i$. Nevertheless, a relationship can be derived between these two concepts, as described in the following proposition. The term added in (6.6) corresponds the cross-entropy between the diagonal joint $\mathcal{P}_{\text{pseud}} = \frac{1}{K}\mathrm{I}_K$ and $\mathcal{P}_{\text{joint}}$, which is bounded as follows:

$$\log K \;\leq\; \mathcal{H}(\frac{1}{K}\mathrm{I}_K, \mathcal{P}_{\text{joint}}) \;\leq\; \frac{1}{N}\sum_{i=1}^{N}\mathcal{H}(\mathbf{u}, \widehat{\mathbf{p}}_i) + \mathcal{H}(\mathbf{u}, \widetilde{\mathbf{p}}_i), \tag{6.7}$$

where $\mathbf{u}$ is the vector such that $u_k = \frac{1}{K}$ for $k = 1, \ldots, K$.

Proof. See Appendix 4. □

Our proposed MI loss can be thus expressed as

$$\mathcal{L}_{\mathrm{MI}} = -I'_{\alpha}(\mathcal{P}_{\mathrm{joint}}) = \mathcal{H}'_{\alpha}(\mathcal{P}_{\mathrm{joint}}) - \mathcal{H}(\widehat{\mathbf{p}}) - \mathcal{H}(\widetilde{\mathbf{p}}) \tag{6.8}$$

As we will show in experiments, purely minimizing the cross-entropy between $\mathcal{P}_{\mathrm{pseud}}$ and $\mathcal{P}_{\mathrm{joint}}$ (i.e., using $\alpha = 1$) does not give optimal results. This is because the true number of clusters is not known, and forcing an arbitrary number of clusters to be balanced is too restrictive. By using a value of α between 0 and 1, as shown in Figure 6.1, enables the network to ignore non-relevant clusters and focus on the most important ones.

6.3.2 Boundary-aware alignment loss for *dense* feature clustering

Clustering dense embeddings based on $\mathcal{L}_{\mathrm{MI}}$ results in balanced and confident clusters, but these clusters do not need to be spatially regular or align with region boundaries in the image. To be useful for the downstream segmentation task, a good representation should capture anatomic structures in the images, whose contours often correspond to regions with strong intensity gradients (i.e., edges). Based on this idea, we propose to use local cross-correlation to match the boundaries of clusters, which correspond to regions with high entropy, with edges in the image. Our cross-correlation loss is defined as follows:

$$\mathcal{L}_{\mathrm{CC}} = \sum_{i\in\Omega} \frac{\left(\sum_{j\in\mathcal{N}(i)} (\phi_j - \hat{\phi}(i)) \cdot (\varphi_j - \hat{\varphi}(i))\right)^2}{\left(\sum_{j\in\mathcal{N}(i)} (\phi_j - \hat{\phi}(i))^2\right) \cdot \left(\sum_{j\in\mathcal{N}(i)} (\varphi_j - \hat{\varphi}(i))^2\right)} \tag{6.9}$$

In this loss, ϕ measures the edge response of a Sobel filter on the input image, while φ is a spatial map of cluster distribution entropy. $\hat{\phi}(i)$ and $\hat{\varphi}(i)$ denote the mean value in a local window $\mathcal{N}(i)$ centered on position i, respectively for ϕ and φ. We note that a similar loss is often used in medical image registration (Balakrishnan, Zhao, Sabuncu, Guttag & Dalca, 2019), where images of two different modalities or acquisitions need to be aligned. Unlike L_2 loss, which imposes a strict equivalence between distributions, this loss can capture correlation in local variance even when images have very different distributions of intensity.

6.3.3 Contrastive loss for *global* feature learning

While the first two losses aim to regularize the local representation of dense feature maps in the decoder, the next one focuses on learning a global representation of the image. Toward this goal, we consider the features produced by the encoder $\phi_{\text{enc}}(x_i)$, which summarize the global context of an input image x_i, and project it into a low-dimensional representation $\mathbf{z}_i$. Similar to (Chaitanya *et al.*, 2020), we regularize global representation $\mathbf{z}_i$ using a contrastive loss exploiting available meta-labels:

$$\mathcal{L}_{\text{con}} = -\frac{1}{2N}\sum_{i=1}^{2N}\frac{1}{|\mathcal{S}_i|}\sum_{j\in\mathcal{S}_i}\log\frac{\exp\left(\mathbf{z}_i^\top\mathbf{z}_j/\tau\right)}{\sum\limits_{a\in\mathcal{S}\backslash\{i\}}\exp\left(\mathbf{z}_i^\top\mathbf{z}_a/\tau\right)}. \tag{6.10}$$

In the loss, $\mathcal{S} = \{i\,|1 \leq i \leq 2N\}$ is the index set of an augmented batch, where each image is randomly transformed twice. Moreover, $\mathcal{G}(i)$ the meta-label of image i and $\mathcal{S}_i = \{j\,|\,\mathcal{G}(j) = \mathcal{G}(i),\ 1 \leq j \leq 2N,\ i \neq j\}$ are the indexes of images within the same meta-label as i. As described in Section 6, we divide volumetric images into different partitions, and use the partition index of each 2D image (slice in the volume) as meta-label. τ is a small temperature factor that helps gradient descent optimization by smoothing the landscape of the loss objective.

Table 6.2 3D DSC on test set when fine-tuned using a few labeled data. Listed methods are applied in a pre-training stage. (*Dec*) means that the loss is applied to dense embeddings in feature maps of the decoder, and (*Enc*) to the global features at the end of the encoder.

Methods	**ACDC-LV**				**ACDC-RV**				**ACDC-Myo**				**PROMISE12**			
	1 scan	**2 scans**	**4 scans**	**mean**	**1 scan**	**2 scans**	**4 scans**	**mean**	**1 scans**	**2 scans**	**4 scans**	**mean**	**4 scans**	**6 scans**	**8 scans**	**mean**
Baseline	67.13	74.49	84.81	75.48	51.82	60.50	64.18	58.84	54.05	67.56	76.00	65.87	49.91	71.53	78.04	66.49
Full Sup.	92.26				86.80				88.07				89.65			
IIC (*Dec*)	71.96	82.84	85.43	80.08	56.92	63.58	64.93	61.81	58.31	70.22	74.98	67.83	54.21	72.97	80.05	69.07
IMSAT (*Dec*)	57.59	75.38	76.76	69.91	34.36	47.81	47.42	43.20	50.52	64.51	71.91	62.23	55.20	74.46	81.50	70.38
Contrast (*Dec*)	64.37	77.69	84.36	75.47	50.75	56.34	50.88	52.66	54.40	70.11	74.05	69.19	54.22	63.52	82.47	66.74
Ours (*Dec*) (only MI)	83.63	86.94	89.33	86.63	66.63	73.78	73.85	71.42	73.65	77.39	81.92	77.08	65.36	78.42	81.92	75.23
Ours (*Dec*) (only CC)	64.85	67.03	79.31	70.70	44.30	50.33	54.52	49.72	49.46	60.13	69.64	59.74	42.48	73.69	80.31	65.50
Ours (*Dec*) (MI+CC)	84.04	**88.52**	89.31	87.29	**76.86**	79.13	75.92	77.30	**76.93**	**79.59**	81.97	**79.49**	68.13	78.75	**82.82**	76.30
Contrast (*Enc*)	80.59	85.68	87.78	84.10	68.91	73.54	72.70	71.72	67.30	77.22	79.58	74.70	63.54	78.24	81.72	74.50
Contrast (*Enc+Dec*)	77.98	85.97	88.42	84.12	66.47	72.82	76.69	71.99	64.96	76.98	78.76	73.57	60.68	77.97	80.53	73.06
Contrast (*Enc*)+ Ours (*Dec*)	**84.48**	87.85	**90.04**	**87.45**	75.42	**79.73**	**78.89**	**78.01**	74.30	78.43	**82.82**	78.52	**69.76**	**80.47**	82.09	**77.44**

6.3.4 Our unified pre-training objective

Our final objective for unsupervised representation learning combines all three objectives as follows:

$$\mathcal{L}_{\text{total}} = \underbrace{\mathcal{L}_{\text{con}}}_{\text{global embedding}} + \underbrace{\lambda\mathcal{L}_{\text{CC}} + \mathcal{L}_{\text{MI}}}_{\text{dense features}} \tag{6.11}$$

$\mathcal{L}_{\text{con}}$ is applied on the global representation of the encoder, and therefore it influences only the encoder. Conversely, $\mathcal{L}_{\text{MI}}$ and $\mathcal{L}_{\text{CC}}$ are used on the dense features of the penultimate layer of the decoder, hence they affect the parameters of the whole network. These three losses are learned jointly in a single pre-training step. See Fig. IV-1 in the Appendix for a graphical illustration.

6.4 Experimental setup

To assess the performance of our proposed pre-training method, we performed extensive experiments on two clinically-relevant segmentation datasets. In this section, we present briefly the experimental setting, employed dataset, as well as implementation details. We include more details in Appendix 5 and 7.

6.4.1 Dataset and evaluation metrics

Two clinically-relevant benchmark dataset are chosen for our experiments: the automatic cardiac diagnosis challenge (ACDC) (Bernard *et al.*, 2018), and the Prostate MR image segmentation 2012 challenge (PROMISE12) (Litjens *et al.*, 2014) dataset. Three foreground classes are delineated for ACDC dataset, which includes left ventricle endocardium (LV), left ventricle myocardium (Myo), right ventricle endocardium (RV), and we consider them as three binary segmentation tasks. Due to the high anisotropic resolution in both datasets, we consider the 2D slices of volumetric images as separate examples, and randomly split them into training, validation and test sets, so that no two images of the same scan are in the same set. To evaluate methods in a setting with limited annotation, we randomly select images from a few scans of the training set as our labeled data set, and consider all the images of the training set as unlabeled. We detail the data pre-processing and augmentation in Appendix 5. For all datasets, we used the 3D Dice similarity coefficient (DSC), which measures the overlap between the predicted labels S and the corresponding ground truth labels G: $\mathrm{DSC}(S, G) = \frac{2\times|S\cap G|}{|S|+|G|}$. In all experiments, we reconstruct the 3D segmentation for each scan by aggregating the predictions made for 2D slices and report the 3D DSC metric for the test set corresponding to the best-performing epoch on the validation set.

6.4.2 Comparable methods and ablation variants

We compare our proposed method with clustering-based and contrastive-based self-supervised learning approaches, as well as six state-of-the-art semi-supervised segmentation methods. IMSAT (Hu, Miyato, Tokui, Matsumoto & Sugiyama, 2017a) and IIC (Ji *et al.*, 2019) also employ MI maximization as the optimization criterion and cluster the local embeddings pixel-wisely. The contrastive learning method relies on the construction of positive and negative pairs. For dense embedding, positive pairs are embeddings in the same position undergoing different intensity transformations, where negative pairs are defined as embeddings with sufficient large distances. Six semi-supervised segmentation methods are also tested, and we present the details of each method in Appendix 6. Lastly, we boost the best performing semi-supervised method with pre-trained weights from our pre-training methods.

6.4.3 Implementation details

We employ U-Net (Ronneberger *et al.*, 2015) as our segmentation network architecture, which consists of five symmetric encoder/decoder blocks with skip connections. We extract the local embeddings in the decoder layer before the last 1×1 convolution, thus they have the same spatial resolution as the input image. These embeddings are then projected to probabilities over K cluster using a projector comprised of a 1×1 convolution and a K-way softmax. We fix $K = 40$ for all datasets. Hyper-parameter α is introduced in our method and we fixed it to 0.5 for all experiments. Image transformation $\mathcal{T}(\cdot)$ consists of gamma correction and random affine transformation. Our proposed method follows a two-stage training strategy: *pre-train* for representation learning and *fine-tune* for evaluation this representation on the downstream segmentation task. In the *pre-train* stage, we optimize the network in an unsupervised way on all training images without pixel-wise annotation, resulting in a set of network parameters θ. We evaluate the quality of these pre-trained weights in a separate *fine-tune* stage by creating a second segmentation network initialized with these parameters, and fine-tuning the whole network using only a few labeled scans. The comparison with other SOTA semi-supervised methods is performed with the same setting as in the *fine-tune* stage, and we report their test DSC performances on their own best hyper-parameters determined by validation performance using grid search. We provide detailed explanation on network architecture, training protocols, transformation $\mathcal{T}(\cdot)$, and hyper-parameters used in each method in Appendix 7.

6.5 Experimental results

In this section, we first compare our method against clustering-based and contrastive-based methods. Then, we evaluate all components of our method as our ablation variants. Finally, we compare our method with the most promising approaches for semantic segmentation in medical imaging, with reduced training data.

6.5.1 Comparison with cluster based methods and ablation variants

Table 6.2 reports the test 3D DSC performance for different representation learning methods on the ACDC and Promise12 datasets. At the top of the table, we report the number of labeled scans used for every result. Reported values are the average over *three* independent runs with different random seeds. Methods presented here all adopt a *pre-train* and *fine-tune* strategy with a few annotated scans.

Upper and lower bounds: We present results for *Baseline*, which uses only the annotated scans with cross-entropy as standard supervised loss, and for *Full Supervision*, where the same loss is used with all available training examples. These represent lower and upper bounds on the expected performance for different methods.

Cluster-based methods: We present in the next two rows the performance for IIC and IMSAT. These two methods employ MI as the optimization objective and perform clustering on local embeddings with K clusters. IIC brings consistent improvements across all four tested classes (4.6%, 3.14%, 1.96%, and 2.58%), while IMSAT leads to a worse performance for the ACDC dataset. We visualize their pre-trained clusters in Fig. 6.2, showing that these methods fail to find balanced clusters corresponding to meaningful regions of the image.

Contrastive-based method: We then report in the next row the performance obtained using contrastive learning only on dense features of the decoder. Surprisingly, we observe that optimizing the contrastive objective with grid-based positive and negative pairs provides no benefit for the segmentation tasks. This is due to very weak guidance offered by contrasting dense embeddings.

Our ablations: We then present in the next three lines the performance for our proposed ablation variants. Our modified $\mathcal{L}_{\text{MI}}$ alone leads to substantial improvements compared to the original IIC: 6.54%, 9.61%, 9.25%, and 6.16% are observed for the four classes. These improvements clearly indicate the advantage of introducing a pseudo-mask $\mathcal{P}_{\text{pseud}}$ to guide the learning of the joint probability matrix. $\mathcal{L}_{\text{CC}}$ aligns cluster boundaries with image edges, but does not help segmentation on its own since predicted clusters are not consistent across images and transformations. Last, we observe that combining our proposed $\mathcal{L}_{\text{MI}}$ and $\mathcal{L}_{\text{CC}}$ lead to significant improvements over using $\mathcal{L}_{\text{MI}}$ alone. These improvements are particularly notable for RV and Myo classes, which are more complex and rely more on image edges.

Global feature pre-training: The last three rows report the performance of methods employing contrastive learning on global features. The method Contrast (Enc) which only optimizes $\mathcal{L}_{\text{con}}$ significantly improves the segmentation quality given a few labeled scans. However, these improvements are still inferior to the Ours (Dec) (MI+CC) variant which, unlike Contrast (Enc), does not use meta-labels. Contrast (Enc+Dec), which combines *global* and *local* contrastive objectives, leads to marginal improvements. Our proposed method is complementary to the *global* contrastive based method. We report in the last row the performance of our proposed method combining all three losses: $\mathcal{L}_{\text{con}}$, $\mathcal{L}_{\text{MI}}$ and $\mathcal{L}_{\text{CC}}$. This method achieves the highest accuracy on 10 out of 16 cases, and second rank for remaining cases. Further, it yields average DSC improvement over Baseline as large as 17.35%, 23.60%, 20.25% and 17.85%, for the LV, RV, Myo and Prostate tasks, respectively.

Table 6.3 Impact of α for our proposed $\mathcal{L}_{\mathrm{MI}}$.

α	ACDC-LV			ACDC-RV			ACDC-Myo		
	1 scan	**2 scans**	**4 scans**	**1 scan**	**2 scans**	**4 scans**	**1 scans**	**2 scans**	**4 scans**
0.0	61.58	78.09	81.27	31.95	23.50	41.64	54.74	58.15	69.45
0.25	**84.36**	87.68	**89.32**	38.98	59.73	59.39	**76.93**	79.59	81.97
0.5	84.04	**88.52**	89.31	**76.86**	79.13	**75.92**	76.27	**79.81**	**82.36**
0.75	82.02	87.81	89.03	76.76	**79.41**	75.49	73.14	78.79	81.79
1.0	81.31	85.58	88.66	73.34	76.37	70.44	71.90	79.47	81.22

Table 6.4 Impact of our proposed boundary-aware loss $\mathcal{L}_{\mathrm{CC}}$.

λ_{CC}	ACDC-LV			ACDC-RV			ACDC-Myo		
	1 scan	**2 scans**	**4 scans**	**1 scan**	**2 scans**	**4 scans**	**1 scans**	**2 scans**	**4 scans**
0.0	83.63	86.94	89.33	66.63	73.78	73.85	73.65	77.39	81.92
0.1	80.54	85.40	87.80	66.99	76.00	75.33	72.62	77.03	81.39
1.0	**84.04**	**88.52**	**89.31**	**76.86**	**79.13**	**75.92**	**76.93**	**79.59**	**81.97**
4.0	78.36	84.60	85.48	70.84	75.39	70.88	73.74	76.65	81.33

6.5.2 Visualization of pre-trained cluster assignments

To better understand our boundary-aware information maximization method, we visualize in Fig. 6.2 different cluster assignments obtained by our ablation variants and compared methods. Clusters obtained at the end of the unsupervised pre-training are illustrated by different colors.

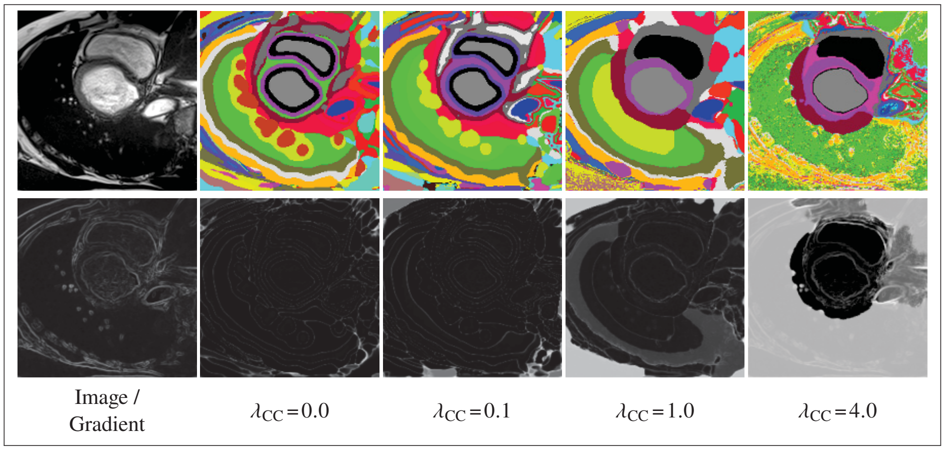


Image /
Gradient
$\lambda_{\mathrm{CC}}=0.0$
$\lambda_{\mathrm{CC}}=0.1$
$\lambda_{\mathrm{CC}}=1.0$
$\lambda_{\mathrm{CC}}=4.0$

Figure 6.3 Boundary loss effect. Upper: Input image and different pre-trained clusters; Down: Image edges and entropy map for each cluster. Black color refers to certain cluster regions while bright regions reflect uncertain predictions.

Table 6.5 Impact of number of clusters K.

K	ACDC-LV			ACDC-RV			ACDC-Myo		
	1 scan	**2 scans**	**4 scans**	**1 scan**	**2 scans**	**4 scans**	**1 scans**	**2 scans**	**4 scans**
5	66.88	73.95	77.33	49.32	52.10	51.74	38.52	48.18	49.25
10	81.00	86.11	87.73	70.75	74.37	71.78	71.16	77.48	80.14
20	69.72	73.37	74.68	59.63	63.13	54.00	76.03	79.34	82.30
40	84.04	**88.52**	89.31	**76.86**	**79.13**	75.92	**76.27**	**79.81**	82.36
60	**85.25**	88.12	**89.74**	70.28	74.78	**76.39**	75.96	79.09	**82.68**

We also compare these clusters with the SLIC super-pixel algorithm (Achanta *et al.*, 2012) that groups pixels based on both intensity and spatial information. We note that IMSAT, IIC and Ours ($\alpha = 0$) produce highly unbalanced clusters, where a few clusters dominate a large portion of pixels and resulting clusters do not correspond well to anatomical structures of the image. In contrast, our proposed variants with $\alpha \geq 0.25$ clearly capture the main structures in cardiac MR images, without any pixel-wise annotation. In most cases, it is able to successfully separate the LV, RV and Myo classes from the background. Additionally, Promise12 images present less contrast but our method still produces relatively better anatomical structures compared with traditional Super-pixel methods, IIC and IMSAT. This explains the huge improvements brought by our method when fine-tune the network using a few labeled images. Last, we notice that contrastive-based approaches can also boost the segmentation performance. However, they lack the intepretability of clusters provided by our method.

6.5.3 Impact of α in our proposed $\mathcal{L}_{\text{MI}}$ objective

To evaluate the impact of the proposed pseudo-label for the joint distribution matrix, we vary different α based on one of our best performing case and report the results in Table 6.3. It can be seen that increasing α from 0 to 0.25 introduces large improvements for all segmentation tasks, which confirms the poor guidance of the IIC objective. Interestingly, we notice that $\alpha = 1$ does not lead to the best performance. This might be because clustering pixels into $K = 40$ regions of similar sizes breaks the anatomical structures of a given image, and a relatively lower α provides

Table 6.6 We compare the performance of our method with other pre-training approaches and state-of-the-art semi-supervised methods on 3D DSC on test set when fine-tuned using a few labeled data. The bold and underline values indicate the best and the second best performing methods, respectively.

Methods	ACDC-LV				ACDC-RV				ACDC-Myo				PROMISE12			
	1 scan	**2 scans**	**4 scans**	**mean**	**1 scan**	**2 scans**	**4 scans**	**mean**	**1 scans**	**2 scans**	**4 scans**	**mean**	**4 scans**	**6 scans**	**8 scans**	**mean**
Baseline	67.13	74.49	84.81	75.48	51.82	60.50	64.18	58.84	54.05	67.56	76.00	65.87	49.91	71.53	78.04	66.49
Contrast (*Enc+Dec*)	77.98	85.97	88.42	84.12	66.47	72.82	76.69	71.99	64.96	76.98	78.76	73.57	60.68	77.97	80.53	73.06
Ours (*pre-train*)	84.48	87.85	90.04	87.45	75.42	79.73	78.89	78.01	74.30	78.43	82.82	78.52	69.76	80.47	82.09	77.44
Entropy Min.	73.79	80.26	86.84	80.30	56.18	62.09	66.27	61.51	57.23	71.10	76.28	68.20	59.78	76.09	78.98	71.62
MixUp	73.30	76.30	84.42	78.01	61.23	63.60	63.14	62.66	55.74	69.80	73.84	66.46	52.09	75.59	81.11	69.60
Mean Teacher (MT)	83.13	87.02	87.70	85.95	61.61	68.76	67.21	65.86	61.55	75.32	78.42	71.76	84.71	85.97	86.93	85.87
UA-MT	81.08	85.03	87.19	84.43	62.06	67.91	66.64	65.54	59.26	73.68	78.61	70.52	66.16	81.79	84.40	77.45
ICT	76.87	78.41	86.34	80.54	60.31	63.42	68.35	64.03	55.91	71.77	77.90	68.53	63.97	77.92	81.39	74.43
Adv. Train.	75.31	74.85	85.85	78.67	55.29	62.25	64.58	60.71	57.68	70.39	75.94	68.00	71.50	78.63	81.35	77.16
MT + Contrast (*Enc+Dec*)	86.37	89.57	90.40	88.78	75.53	78.42	77.22	77.06	76.11	80.21	82.00	79.44	76.16	82.89	84.85	81.30
MT + Ours (*pre-train*)	**90.25**	**91.36**	**91.04**	**90.88**	**80.16**	**81.50**	**78.97**	**80.21**	**78.71**	**83.33**	**83.61**	**81.88**	**85.64**	**85.60**	**88.45**	**86.56**

a softer guidance that helps preserve these structures. We confirm this by visualizing in Fig. 6.1 the joint matrix $\mathcal{P}_{\text{joint}}$, the cluster assignment, as well as the uncertainty of these clusters for different α.

6.5.4 Impact of boundary-aware loss $\mathcal{L}_{\text{CC}}$

Our boundary-aware loss $\mathcal{L}_{\text{CC}}$ is a key component to boost performance for harder segmentation tasks such as RV and Myo. To determine the usefulness of this loss, similar to the previous experiment, we vary λ_{CC} ranging from 0.0 to 4.0 for our best performing case, and present the results in Table 6.4. Clearly, increasing λ_{CC} from 0.0 to 1.0 improves the segmentation performance for all tasks, in particular for the RV and Myo classes whose boundary mainly follows the image edges. In Fig. 6.3, we show the cluster boundaries separate images obtained with the different λ_{CC}. The boundary-aware loss successfully guides the cluster boundaries towards image edges and reduces the over-segmentation of pixels around the boundaries.

6.5.5 Impact of over-segmented clustering numbers K

Our MI-based method converts continuous feature vectors to a discrete distribution over clusters. The cluster number K is another important hyper-parameter for our method. In this ablation experiment, we measure the impact of K by varying it from 5 to 60. Table 6.5 and Fig. IV-2 in Appendix 8 show the DSC performance and the corresponding cluster assignment obtained with unsupervised pre-training. $K = 5$ leads to a weak segmentation performance, which can be explained by a collapsed cluster assignment. The cluster maps become more balanced with the increase of K and gradually reflect the cardiac structures in the image. However, using $K = 60$ does not give a better performance in segmentation tasks, since the resulting clusters over-segment the image and capture less relevant regions.

6.5.6 Comparison with state-of-the art methods

We compare our method with other approaches that aim to improve training with few annotated images/scans. Table 6.6 presents results for various semi-supervised learning approaches. For a more detailed explanation of the experimental setup of each method, see Appendix 6. To have a fair comparison, for all methods, we used grid search on the validation set to tune the hyper-parameters and report the corresponding test performance. For most methods, the improvement with respect to the baseline trained with only the supervised loss is quite limited and varies depending on the segmentation task and the number of annotated scans used. Among different methods, MT offers stable improvements across all tasks and reaches competitive performance compared with contrast (*Dec+Dec*) and Ours (*pre-train*) for the LV and Prostate classes, mainly due to its temporally-ensembled teacher network which provides stable prediction proposals for unlabeled images. However, semi-supervised methods such as MT are normally trained with randomly initialized parameters and can thus be further improved with our proposed *pre-train* approach as initialization of the network. To test this idea, we ran MT with two different initializations, one from contrastive-based Contrast (*Enc+Dec*) and the other from Ours (*pre-train*). The results in the last two rows of Table 6.6 indicate that a further improved segmentation is obtained by simply initializing the network parameters with these pre-trained checkpoints: Contrast (*Enc+Dec*) boosted the performance of MT by 2.82%, 11.20% and 7.68% for LV, RV, and Myo, while these improvements increase to 4.93%, 14.35% and 10.12% when Ours (*pre-train*) is used as initialization. In summary, our boundary-aware algorithm for unsupervised representation learning can boost state-of-the-art semi-supervised segmentation approaches to achieve excellent segmentation quality, even when only an extremely small amount of labeled examples are available.

6.6 Discussion and conclusion

In this paper, we presented a boundary-aware information maximization method for the unsupervised pre-training of models for medical image segmentation. This method complements the *global* contrastive loss and can highly improve the performance of a segmentation network when annotated data is scarce. It was shown that, with a reduced amount of unlabeled images, our method can learn a useful local representation on dense feature maps during pre-training, without any supervisory signal. Furthermore, as shown in our visualization of results, the clusters obtained by the pre-trained checkpoint enhance intepretability. We compared our method with recent self-supervised learning approaches, based on clustering and contrastive learning, as well as six strong semi-supervised segmentation algorithms. Results on two benchmark datasets demonstrate the outstanding accuracy of our method. In particular, the combination of our method with Mean Teacher yields unprecedented performance, reaching close to full supervision with a single scan.

CONCLUSION AND RECOMMENDATIONS

The Introduction and Background chapters of this thesis highlighted the challenges of learning an accurate segmentation model with reduced supervision signals, such as weak labels and anatomical priors, or with on unlabeled data. In the methodological chapters of the thesis (Chapters 2 – 6), we proposed different techniques to address these challenges, which resulted in four separate contributions.

Our *first* contribution leveraged weak annotations and discrete anatomical priors to improve the learning of a neural network, so that it can achieve a higher segmentation quality and a better constraint satisfaction. The next four contributions focused on the problem of semi-supervised segmentation, where a small amount of labeled data is complemented with a large set of unlabeled data. For this problem, we designed different approaches relying on unlabeled data to guide the network towards a better segmentation. As our *second* contribution, we proposed a co-training based framework leveraging the collaboration between multiple segmentation networks. In this framework, unlabeled images are used to enforce consistency (ensemble agreement) and diversity across predictions, thereby encouraging the exchange of complementary knowledge between the models. As *third* contribution, we then proposed a method based on mutual information (MI) maximization to regularize different feature embeddings in the encoder and decoder of a segmentation network. Specifically, the method maximizes the MI between feature map embeddings obtained from different random augmentations of the same image, which enforces the global invariance and local smoothness of the segmentation. Our *fourth* contribution employed representation learning to boost the segmentation accuracy using unlabeled data. To deal with the noisiness of meta-labels (positive or negative pairs) used for contrastive self-supervised learning, we proposed a self-paced contrastive learning algorithm that focuses on confident pairs to obtain a better representation for the downstream segmentation task. We also showed the benefit of combining different meta-labels on segmentation accuracy. Our *last* contribution, which also focuses on representation learning, proposed an effective clustering-based approach for dense feature pre-training. The proposed method improves the previous MI maximization approach by encouraging the joint cluster probability matrix to be diagonal, and by adding a boundary-aware loss that favors clusters aligned with edges in the images. Experimental results on two clinically-relevant datasets confirmed the effectiveness of our proposed method to boost the segmentation accuracy using as few as one labeled example.

In this section, we give a detailed summary of each contribution, and discuss its practical impact, current limitations and possible directions for future work.

Contribution 1: Discretely-constrained deep network for weakly supervised segmentation

In Chapter 2, we proposed an efficient strategy for weakly-supervised segmentation that imposes regularization priors as discrete constraints on target regions. This method, among the first to propose a discrete optimization strategy for deep segmentation networks, enhances the efficiency and accuracy of the model using a discretely-constrained formulation. Our proposed method is based on the alternating direction method of multipliers (ADMM) algorithm and trains a CNN with discrete constraints and regularization priors. To evaluate its efficiency, we applied our method to the segmentation of medical images with scribbles as the partial annotation, where both size constraints and boundary length regularization are enforced. Results showed our method to provide significant improvements compared to existing approaches in terms of segmentation accuracy, constraint satisfaction and convergence speed.

Impact and limitation:

This method successfully enabled the optimization of the networks with discrete constraints, where network parameters and segmentation proposals for the size and boundary length constraints were alternatively updated with the help of Lagrangian multipliers. A broader set of discrete constraints, such as convexity and connectivity of the shape or the compactness of the segmentation region could also be investigated. On the other hand, our method requires to provide discrete proposals throughout the training, which can increase the training time and computational cost. A potential solution for this problem is to create a proposal bank per constraint and updating these proposals before each epoch starts. Parallelizing the generation of these discrete proposals on multiple CPU cores could also help reduce the computational time and thus improve the overall training efficiency.

Contribution 2: Deep co-training for semi-supervised image segmentation

In Chapter 3, we focused on semi-supervised segmentation and proposed a co-training based algorithm for this problem. The success of this algorithm relies on two losses, the first one enforcing consistency (agreement) for different model predictions given unlabeled images, and the second one boosting the prediction diversity across models with adversarial images. This method has shown to be effective on three challenging medical image segmentation tasks covering different modalities, where it boosted segmentation accuracy while using few labeled images. To better understand the proposed method, we visualized these adversarial images generated by the diversity. We also evaluated the benefit of using more than two models and showed it can bring a higher improvement but with an increased computational cost.

Impact and limitation:

The proposed method provides a generic way to use co-training for semi-supervised segmentation, without the need to define different views for the separate models. However, since this method uses multiple models during both training and inference, it can lead to slow computation and large memory usage, especially for segmenting volumetric data such as CT and MRI images. Distilling the knowledge from the ensemble of trained models to a small network could be explored to reduce computational and memory requirements for model deployment. In the results, we provided the average and standard derivation of performance over multiple runs for each experiment. However, a statistical test is also recommended to confirm the significance of improvements provided by our method. Moreover, since the proposed method is orthogonal to other semi-supervised learning approaches such as Mean Teacher (Tarvainen & Valpola, 2017) and Pseudo-Labels (Lee *et al.*, 2013), another direction of research to further boost the accuracy of our method would be to combine it with these approaches.

Contribution 3: Boosting semi-supervised image segmentation with global and local mutual information regularization

Chapter 4 presented a novel semi-supervised segmentation method using the mutual information (MI) on categorical distributions to achieve both global representation invariance and local smoothness of features maps. The method applied a global MI loss on the embeddings from the encoder to enforce transformation invariance. A local MI loss was also used to promote spatial consistency on dense feature embeddings from intermediate layers of the decoder, leading to a smoother segmentation. Since MI is difficult to measure for high-dimensional continuous distributions, we employed projection heads to convert feature maps to discrete distributions representing cluster assignments, where MI can be computed efficiently. This method has been evaluated extensively on four challenging publicly-available datasets for medical image segmentation. Results showed that our method outperforms recently-proposed approaches for semi-supervised segmentation and provides an accuracy near to full supervision when training with very few annotated images.

Impact and limitation:

While the proposed approach was used on an encoder-decoder based segmentation network, it could be easily extended to other, more recent architectures such as Vision Transformers (Valanarasu, Oza, Hacihaliloglu & Patel, 2021). Our method is also orthogonal to existing semi-supervised methods and, as shown in the ablation study, can be combined with techniques like Mean Teacher (Tarvainen & Valpola, 2017) to improve performance in low-data regimes.

However, the work related to this contribution has some limitations that could be addressed in future research. First, while our method only used random flips as transformation, considering a broader set of transformations could lead to a better performance. Second, other approaches to estimate MI could be investigated. In our method, MI is maximized by converting continuous variables to discrete distributions via a linear projector. Comparing this approach to recent MI estimation techniques, such as InfoNCE (Oord *et al.*, 2018) and MINE (Belghazi *et al.*, 2018), could further demonstrate the usefulness of MI maximization for semi-supervised learning. Last, whereas our method was tested on 2D images, it could also be employed to segment volumetric images using a 3D CNN.

Contribution 4: Self-paced contrastive learning for semi-supervised medical image segmentation with meta-labels

Chapter 5 aimed at acquiring a useful representation from unlabeled images. Specifically, we adapted contrastive learning to train the encoder of a segmentation network using meta-labels stating if a pair of 2D slices come from the same location of their corresponding MRI scan, are from the same subject, or were acquired at the same instant of the cardiac cycle. To mitigate the potential noise in these meta-labels, we proposed a self-paced learning strategy that focuses on confident meta-labels in earlier stages of training. We verified the effectiveness of the proposed method on five medical image segmentation datasets. Results showed that the proposed self-paced mechanism can improve contrastive learning in a noisy context, and boost segmentation quality for down-stream tasks with very few labeled examples.

Impact and limitation: This work revealed the potential of self-paced contrastive learning for image segmentation and the usefulness of combining different types of easy-to-obtain meta-labels. Our best results were obtained by combining discriminative pre-training with semi-supervised learning, showing for the first time the complementary nature of these two training paradigms. Our segmentation method was tested on 2D slices with meta-labels generated from volumetric images. Finding useful meta-labels to train a 3D segmentation network could be investigated in future work. In addition, our proposed method has multiple hyper-parameters to tune, such as the choice of self-paced regularizer (hard or linear), the weighting coefficient for each loss objective, as well as the lower and upper bounds for the self-paced ramp-up scheduler. Another line of research could be to develop strategies to automatically adjust these hyper-parameters (Neary, 2018).

Contribution 5: Boundary-aware information maximization for self-supervised medical image segmentation

In Chapter 6, we presented a clustering-based representation learning approach for dense feature map pre-training. This approach exploits a mutual information maximization objective which groups dense features into balanced and confident clusters. A boundary-aware regularization loss is also used to align the cluster boundaries, which have a higher clustering entropy, with edges in the image. Our method operating on dense representations extends previous global contrastive learning techniques that only optimize the encoder of a segmentation network. Experimental results on two clinically-relevant segmentation benchmarks revealed the effectiveness of our proposed method. Results showed our method to outperform different contrastive-based pre-training methods as well as several state-of-the-art semi-supervised approaches, given a few densely-annotated examples.

Impact and limitation: Contrastive approaches, which heavily relies on the quality of positive and negative pairs, often results in limited improvements for dense feature map pre-training. In contrast, our boundary-aware information maximization method learns a useful local representation on dense feature maps without the need to define these pairs. Furthermore, as illustrated in our visualization analyses, our method helps interpret the learned representation since the clusters on this representation often correspond to anatomical structures in the images to segment. This visual interpretability can not be easily achieved by other contrastive-based approaches. While our method achieved substantial improvements on the task cardiac MRI segmentation, it requires images with a sufficient contrast between the regions of interest in order to extract edge information. Further experiments, testing the proposed method on a broader range of data with different appearance and acquisition protocols, are recommended as future work.

APPENDIX I

INFORMATION-BASED CLUSTERING: AN EXPERIMENTAL STUDY

Jizong Peng[1] , Christian Desrosiers[1] , Marco Pedersoli[2]

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

Article available on *Axiv*[1] in Oct. 2019

1. Presentation

We present in this appendix our work entitled “Information-based Clustering: an experimental study", which focuses on deep clustering based on mutual information maximization. In this work, we present a generalization of information-based deep clustering where two key factors are evaluated: i) the variables for which we want to maximize the mutual information, ii) the regularization of the mutual information loss by the use of image transformations. Through an extensive analysis, we show that maximizing the mutual information between a sample and its transformed version, with an additional regularization to make the learning smoother, outperforms previous approaches and leads to state of the art results on three different datasets. Additional experiments show that the proposed method largely outperforms disentangling methods for classification tasks and is useful as unsupervised initialization for supervised learning

2. Introduction

In deep learning, supervised methods have shown excellent performance, sometimes even surpassing human level (Krizhevsky *et al.*, 2012; He, Zhang, Ren & Sun, 2015). However, these

[1] https://arxiv.org/abs/1910.01665

methods require large datasets with fully annotated data, which cannot be afforded in many cases (Vondrick, Patterson & Ramanan, 2013). Instead, unsupervised methods can learn from data without annotations, which is very appealing given the large amount of data that can easily be collected from various sources such as social media (Schroff, Criminisi & Zisserman, 2011). In this paper, we are interested in the unsupervised learning problem of deep clustering, which consists in learning to group data into clusters, while at the same time finding the representation that best explains the data. Jointly learning to group data (clustering) and represent data (representation learning) is an ill-posed problem which can lead to poor or degenerate solutions (Yang, Parikh & Batra, 2016; Xie, Girshick & Farhadi, 2016; Caron *et al.*, 2018). A principled way to avoid most of these problems is mutual information (Paninski, 2003). Mutual information is a powerful approach for clustering because it does not make assumptions about the data distribution and reduces the problems of mode collapse, where most of the data is grouped in a single large cluster (Caron *et al.*, 2018).

In recent publications, two papers obtained outstanding results for deep clustering by using mutual information in different ways. The first one, IMSAT (Hu *et al.*, 2017b) maximizes the mutual information between input data and the cluster assignment, and regularizes it with virtual adversarial samples (Miyato *et al.*, 2018) by imposing that the original sample and the adversarial sample should have similar cluster assignment probability distribution (by minimizing their KL divergence). The second one, IIC (Ji *et al.*, 2019), maximizes the mutual information of the cluster assignment between a sample and the same sample after applying a geometrical transformation and no additional regularisation is used.

In this paper, we aim to analyze and better understand these algorithms by decomposing them in two basic building blocks: the information-based loss and a regularization term based on image transformations. We build a generalization of information-based clustering approaches in which IMSAT and IIC are special cases. We consider the two different ways to use mutual information for clustering and we evaluate the effect of a regularization based on different image transformations currently used for data augmentation such as Geometrical transformations, Gaussian noise, Virtual Adversarial Training (VAT) (Miyato *et al.*, 2018), Mixup (Zhang *et al.*, 2017b), Cutout (Devries & Taylor, 2017). Form our extensive evaluation on three different datasets we found that: *i*) maximizing the mutual information between a cluster distribution and its transformed version seems to be more robust than other approaches when dealing with challenging datasets; *ii*) adding a transformation-based regularization to the mutual information loss can make the training smoother and leads to better clusters; *iii*) using geometrical transformations for the mutual information loss together with a regularization based on VAT (Miyato *et al.*, 2018) leads to improved results in most of the datasets. Additionally, our best model outperforms popular methods for disentangling representations and can be used as initialization to improve supervised training.

In the reminder of this paper, we first introduce related work in section 3, with special attention to the two mentioned methods. Then, in section 4, we propose the experimental protocol by defining the different components of our experiments. Finally, we report results in section 5 and draw conclusions in section 7.

3. Related work

Mutual information Mutual information $I(X, Y)$ is a information-theoretic criterion to measure the dependency between two random variables X, Y (Bridle *et al.*, 1992). It is defined as the KL divergence between the joint distribution $p(X, Y)$ of two variables and the product of their marginal: $I(X; Y) = D_{\mathrm{KL}}(p(X, Y), p(X)p(Y))$. The criterion of maximizing mutual information for clustering is first introduced in Bridle *et al.* (1992), as the *firm but fair* criterion. In this case the mutual information between input data (i.e, image or representation) and output categorical distribution is maximized, believing that the class distribution can be deduced given the input. This principle is extended in Krause *et al.* (2010), in which mutual information is maximized with additionally an explicit regularization, such as L_2 loss. This helps to avoid too complex decision boundaries.

In Viola & Wells III (1997); Pluim, Maintz & Viergever (2003); Thévenaz & Unser (2000), mutual information is used to align inputs with different modalities (Viola & Wells III, 1997; Pluim *et al.*, 2003; Thévenaz & Unser, 2000), since with different modalities normal distances are meaningless. Finally, mutual information is also used as regularization in a semi-supervised setting (Manohar, Povey & Khudanpur, 2015). Recently, DeepInfoMax (Hjelm *et al.*, 2018) simultaneously estimates and maximizes mutual information between images and learns high-level representations. However, estimating the mutual information of images is challenging and requires complex techniques (Belghazi *et al.*, 2018). Finally, two recent techniques for deep clustering based on mutual information are IMSAT (Hu *et al.*, 2017b) and IIC (Ji *et al.*, 2019). As these are the starting point of this study, they will be analyzed in more detail in section 4.

Self-supervised approaches Self-supervised learning has recently emerged as a way to learn a representative knowledge based on non-annotated data. The main principle is to transfer the unsupervised task to a supervised one by defining some *pseudo labels* that are automatically generated by a *pretext task* without involving any human annotations (Jing & Tian, 2020). The network trained with the pretext task is then used as the initialization of some downstream tasks, such as image classification, semantic segmentation and object detection. It has been shown that a good pretext task can help improve the performance of the downstream task (Pathak, Agrawal, Efros & Darrell, 2017; Doersch & Zisserman, 2017; Noroozi *et al.*, 2018; Caron *et al.*, 2018; Ahsan, Madhok & Essa, 2019). Without the access to label information, the pretext task is usually defined based on the data structure and is somewhat related but different from the downstream tasks. Various pretexts have been proposed and investigated. Similarity-based methods (Noroozi *et al.*, 2018; Caron *et al.*, 2018) design a pretext task that let the network learn the semantic similarities between image patches. Likewise, Noroozi & Favaro (2016); Ahsan *et al.* (2019); Wei *et al.* (2019) trained a network to recognize the spatial relationship between image patches. See Jing & Tian (2020) for a complete survey on such methods.

In this work, we argue that if the task we want to learn with non-annotated data is classification then the best pretext task is clustering. With clustering, the pretext task is very close to the downstream task of classification. In fact, clustering aims to group the data in a meaningful way and therefore split the data into categories. If these categories are not only visually similar but also semantically, classification and clustering become the same task. In other words, with clustering, there is not need for a downstream task. By assigning the most likely category to each cluster, our clustering method becomes a classifier. In the experimental evaluation, we will compare the performance of our information based clustering with state-of-the-art self-supervised learning approaches. For doing that, we use two simple assumptions: i) the sample distribution per class is known (normally uniform) and ii) the exact number of classes, which corresponds to the number of clusters is also known.

Clustering approaches Clustering has been long time studied before the deep learning era. K-means (Kanungo *et al.*, 2002) and GMM algorithms (Banfield & Raftery, 1993) were popular choices given representative features. Recently, much progress has been made by jointly training a network that perform feature extraction together with a clustering loss (Xie *et al.*, 2016; Li, Qiao & Zhang, 2018a; Ghasedi Dizaji, Herandi, Deng, Cai & Huang, 2017; Caron *et al.*, 2018). Deep Embedded Clustering (DEC) (Xie *et al.*, 2016) is a representative method that uses an auto-encoder as the network architecture and a cluster-assignment hardening loss for regularization. Li et al. (Li *et al.*, 2018a) proposed a similar network architecture but with a boosted discrimination module to gradually enforce cluster purity. DEPICT (Ghasedi Dizaji *et al.*, 2017) improved the clustering algorithm's scalability by explicitly leveraging class distributions as prior information. DeepCluster (Caron *et al.*, 2018) is an end-to-end algorithm that jointly trains a Convnet with K-means and groups high-level features to N pseudo labels. Those pseudo labels are in turn used to retrain the network after each iteration. In this work, we focus on two state of the art information-based clustering approaches (Hu *et al.*, 2017b; Ji *et al.*, 2019) and analyze their different components and how they can be combined in a meaningful way.

4. Information-based clustering

We consider information-based clustering as a family of methods having two main components: the mutual information loss and possibly a regularisation based on image transformations. The maximization of the mutual information aims to produce meaningful groups of data, i.e. clusters with a similar representation and with a even number of samples. On the other hand, transformations are used to make the learned representation locally smooth and ease the optimization in a similar way as in data augmentation. For these two components, we consider and evaluate different possible choices and their combination.

4.1 Mutual Information losses

$\mathbf{I(Y, \widetilde{Y})}$: This formulation introduced in IIC (Ji *et al.*, 2019) maximizes the mutual information between the clustering assignment variable Y and the clustering assignment of a transformed sample $\widetilde{Y}$. Mutual information is defined as the KL divergence between the joint probability of two variables and the product of their marginals (Bridle *et al.*, 1992):

$$I(X_1, X_2) = D_{\mathrm{KL}}(p(X_1, X_2),\ p(X_1)\, p(X_2)). \tag{A I-1}$$

It represents a measure of the information between the two variables. If the variables are independent, the mutual information is zero because the joint will be equal to the product of the marginals. Thus, we need to estimate the joint probability of the clustering assignment and its transformed version $p(Y, \widetilde{Y})$, as well as the marginals $p(Y)$ and $p(\widetilde{Y})$. The marginals are defined as

$$p(Y) = \mathbb{E}_X[p(\,Y|X)\,], \quad p(\widetilde{Y}) = \mathbb{E}_X[\,p(Y|T(X))\,], \tag{A I-2}$$

and can be empirically estimated by averaging output mini-batches. For the joint, we compute the dot product between $Y|X$ and $Y|T(X)$ for each sample and marginalize over X:

$$p(Y,\widetilde{Y}) \;= \mathbb{E}_X[p(Y|X) \;\cdot\; p(Y|T(X))^\top]. \tag{A I-3}$$

For each sample, the joint probability of $c_1 \in Y|X$ and $c_2 \in Y|T(X)$ is $c_1\,c_2$. For a single sample, by construction the joint and the marginals will be equivalent. However, when marginalizing the joint over samples X, the final $p(Y,\widetilde{Y})$ will be different than $p(Y)p(\widetilde{Y})$.

This formulation maximizes the predictability of a variable given the other. It is different than enforcing KL divergence between two distributions because: *i*) it does not enforce the two distributions to be the same, but only to contain the same information. For instance, one distribution can be transformed by an invertible operation without altering the mutual information. *ii*) it penalizes distributions that do not have uniform marginal, i.e. all the cluster should contain an even number of samples.

I(X, Y): We consider the MI formulation used as loss in IMSAT (Hu *et al.*, 2017b). It connects the input image distribution X with the output cluster assignment of the used neural network Y and is the most common way to use mutual information for clustering (Bridle *et al.*, 1992; Krause *et al.*, 2010). As the input of the neural network is a continuous vector, estimating its probability distribution is hard and we cannot use directly equ. (A I-1). Instead, in IMSAT the mutual information between input and output is computed as:

$$I(X,Y) \;=\; H(Y) - H(Y|X) \tag{A I-4}$$

$$= \mathbb{E}_Y[\,\log \mathbb{E}_X[\,p(Y|X)\,]\,] - \mathbb{E}_{X,Y}[\,\log p(Y|X)\,] \tag{A I-5}$$

In this formulation, the mutual information is easy to compute because it is the difference between the entropy of the output $H(Y)$ and the conditional entropy $H(Y|X)$. Both quantities can be approximated in a mini-batch stochastic gradient descent setting. $H(Y)$ is approximated as the entropy of the average probability distribution $p(Y|X)$ over the given samples, while $H(Y|X)$ is approximated as average of the conditional entropy of each sample. As entropy is a non-linear operation, the two quantities are different. $H(Y)$ is maximized when the probability of each cluster is the same, i.e. the output has the same probability distribution for each cluster. On the other hand, $H(Y|X)$ is minimized when, for each sample X, $p(Y|X)$ has most of the probability distribution assigned to a single cluster, i.e. the model is certain about a given choice. Combining the two imposes that the clustering chooses a single cluster for every sample and, globally, each cluster contains the same number of samples. In case the class distribution is not uniform, another distribution C can be enforced by $D_{\mathrm{KL}}(p(X), p(C))$. In our experiments, we limit ourselves to uniform class distribution. Notice that, in IMSAT, the authors add an hyper-parameter MI formulation $\lambda H(Y) - H(Y|X)$. This parameter is the minimum value that ensures that the data is evenly distributed on all clusters. We follow the same approach as in the original paper.

4.2 Regularization:

In the context of this work, we call regularization the additional loss that penalizes when the output of the model for the original image and the transformed image are different. Although in unsupervised settings there is no real difference between loss and regularization because in both cases the training is performed without annotations, here we consider this KL divergence as a regularization because it cannot be used alone for clustering. Another term such as $H(Y)$ is required to enforce an even distribution of samples in the clusters.

As we have access to the clustering probability distribution $p(Y|X)$ of a sample X and its transformed version $p(Y|T(X))$, we use the KL divergence as penalty term $D_{\mathrm{KL}}(p(Y|X), p(Y|T(X)))$.

While for $I(X,Y)$ this regularization is fundamental for good results (Krause *et al.*, 2010; Hu *et al.*, 2017b), for $I(Y,\widetilde{Y})$ the original paper did not use any additional regularization. In this work, we analyze the effect of regularization with different transformations for both approaches. The following section describes in more detail the used transformations.

4.3 Transformations

Transformations seem to be a key component of MI-based clustering. To be useful, any transformation needs to change the appearance of the image (in terms of pixels) while maintaining its semantic content, i.e. the class of the image. We can think of these transformations as a pseudo ground-truth that helps to train the model. In this work, we consider five types of transformations: Geometrical, Gaussian, Adversarial, Mixup and Cutout.

Geometrical: Geometrical transformations are the image transformations that are normally used for data augmentation. As in (Ji *et al.*, 2019), we use random crop, resize at multiple scales, horizontal flip, and color jitter. Note that some transformations can actually change the category of a class. For instance, on MNIST (LeCun, Bottou, Bengio, Haffner *et al.*, 1998) a dataset composed of numbers, a crop of a 6 can zoom in the lower circle and look very similar to a 0. We will further discuss this problem in the experimental results.

Gaussian: Adding Gaussian noise to the input space, as a regularization method, has been proposed in (Rifai, Glorot, Bengio & Vincent, 2011). It has been demonstrated that noise injection can effectively reduce the L2-norm of the Jacobian's mapping function with respect to the input (Rifai *et al.*, 2011), thus implicitly enforce regularization similar to L2 constrains. In this paper, we consider to add Gaussian independent and identically distributied noise having 0 mean and θ_{std} standard derivation to each pixel of the three channels of the original input image x:

$$x_{\text{Gauss}} = x + \mathcal{N}(0, \theta_{\text{std}}). \tag{A I-6}$$

Adversarial: Adversarial samples (Yuan, He, Zhu & Li, 2019) are samples that are slightly modified by an adversarial noise which is usually unnoticeable by the human eye, but can induce a neural network to misclassify an example. Recently, methods based on adversarial examples have attracted a lot of attention because they can easily fool machine learning algorithms and thus represent a threat to any system using machine learning (Su, Vargas & Sakurai, 2019; Chou, Tramèr, Pellegrino & Boneh, 2018). It has been shown (Madry, Makelov, Schmidt, Tsipras & Vladu, 2017) that adding those samples during training can help to improve the robustness of the classifier. In this study, we use Virtual Adversarial Training(VAT) (Miyato *et al.*, 2018), an extension of adversarial attack that can also be used for non-labelled samples and has shown promising results for fully-supervised, semi-supervised (Miyato *et al.*, 2018) and unsupervised learning (Hu *et al.*, 2017b). The adversarial noise r can be found as the value within a certain neighbourhood ϵ that maximizes the distance D between the probability distribution of the original sample x and the transformed sample $x + r$:

$$r_{\mathrm{vadv}} = \underset{\|r\|_2 \leq \epsilon}{\arg\max}\ D\left[p(y|x,\theta),\ p(y|x+r)\right], \tag{A I-7}$$

where $D[p_1(.), p_2(.)]$ is a divergence function, usually defined as KL-divergence. In practice, equ. (A I-7) can be optimized in order to find r with a few iterations of the power method (Journée, Nesterov, Richtárik & Sepulchre, 2010). Note that we could also experiment with adversarial geometrical transformations as in (Peng, Tang, Yang, Feris & Metaxas, 2018), but we leave this direction as future work.

Mixup: This is a simple data augmentation technique that has proven successful for supervised learning (Zhang *et al.*, 2017b). It consists on creating a new sample x_{new} and label y_{new} by linearly combining two training samples x_1 and x_2 (e.g. images) and labels y_1 and y_2 (e.g. class probabilities):

$$\begin{aligned} x_{\text{new}} &= \alpha x_1 + (1-\alpha)x_2 \\ y_{\text{new}} &= \alpha y_1 + (1-\alpha)y_2. \end{aligned} \tag{A I-8}$$

α is the mixing coefficient and is normally sampled from a β distribution. Although very simple and effective, Mixup has received multiple criticisms because it is clear that the generated images do not represent real samples. However, the β distribution has most of its mass near 0 and 1, which means that in most of the cases the mixed samples look very similar to one of the samples, but with a structured noise coming from the other image. This transformation differs from the previous one because it requires two input samples to generate a new one. Thus, to use it in our family of algorithms, we had to adapt it. For $I(Y, \widetilde{Y})$, as before we consider $Y = \mathbb{E}_X[p(Y|X)]$ the expected output of real samples X, while $\widetilde{Y} = \mathbb{E}_{X,X_2,\alpha}[\alpha p(Y|X)(1-\alpha)P(Y|X_2)]$ is now the output associated to mixup samples generated using the same real samples X in combination with other samples X_2 that is randomly selected. Mixup can also be used as direct regularization (see next section). In this case, the first output distribution is associated to real samples while the second is associated to mixup samples built as above.

Cutout: Cutout (Devries & Taylor, 2017) is an effective data augmentation method to boost the generalization of a neural network, and has been used in (self)-supervised learning (Pathak, Krahenbuhl, Donahue, Darrell & Efros, 2016), and weakly supervised localization (Singh & Lee, 2017). Cutout forces the network to consider the global information of an image by randomly masking some regions of it. We found this data augmentation suitable for clustering, in the sense that a masked image should have very similar cluster category as the original image.

5. Experimental Setup

Our main experiment evaluates on the three datasets (presented below) the two identified components for information based clustering: information based losses and regularization based on image transformations. For completeness, we have reported all results with all combinations of the different components in the supplementary material.

5.1 Datasets

We evaluate the different methods on 3 datasets:

- MNIST dataset (LeCun *et al.*, 1998) of hand-written digit classification consists of 60,000 training images and 10,000 validation images. 10 classes are evenly distributed in both train and test sets. Following common practice, we mix the training and test set to form a large training set. During training, we do not show any ground truth information, while for testing, we use image annotations to find a mapping between true class label and cluster assignment, thus assessing the clustering performance by the classification accuracy.
- CIFAR10 (Krizhevsky, Hinton *et al.*, 2009) is a popular dataset consisting of 60,000 32×32 color images in 10 classes, with 6,000 images per class. Similar to MNIST dataset, we mix the 50,000 training images with 10,000 test images to build a larger dataset for clustering.
- SVHN (Netzer *et al.*, 2011) is a real-world image dataset for digit recognition, consisting of 73,257 digits for training, 26,032 digits for testing. Images come from natural scene images. We adopt the previously described strategy to use this dataset too.

5.2 Evaluation Metric

Our method groups samples into clusters. If the grouping is meaningful, it should be related to the datsest classes. Thus, in most of our experiments, we use classification accuracy as measure of the clustering quality. This makes sense because the final aim of this approach is exactly to produce a classifier without using training labels. This accuracy is based on the best possible one-to-one mapping (using the Hungarian method (Kuhn & Yaw, 1955)) between clustering assignment and ground truth label (assuming they share the same number of classes). We run the experiments 3 times with different initialization and report mean and standard deviation values.

5.3 Implementation Details

In order to provide a fair comparison, we use the same network for a given dataset across methods. For both MNIST and SVHN dataset, we borrow the setting of IIC (Ji *et al.*, 2019), using a VGG-based convolutional network as our backbone network. For CIFAR-10, we use a ResNet-34 (He, Zhang, Ren & Sun, 2016) based network. It is worth mentioning that, in original IMSAT paper (Hu *et al.*, 2017b), the used network was just a 2 fully connected layers with pre-trained features on CIFAR-10 or GIST features (Oliva & Torralba, 2001) on SVHN. Instead, in this work, we want to compare all results on the same convolutional architecture and without pre-trained models or any hand-crafted features.

For our best method, as in Ji *et al.* (2019) we use two additional procedures to further improve results. The first one, over-clustering, consist in using more clusters than the number of classes in the training data. This can help to find sub-classes and therefore reduce the intra-class variability on each cluster. The second consists in splitting the last layer of the network in multiple final layers (there called heads) and therefore multiple clusters. This can increase diversity and acts as a simplified form of ensembling. Combining these two techniques can highly boost the final performance of the clustering approach. However, they also increase the computational cost of the model. Thus, for the evaluation of all configurations in a same setting, we use a basic model without additional over-clustering or multiple final layers. However, for our best configuration, we retrained it with 5 final layers with 10 clusters (as the number of classes) and other 5 final layers with 50 clusters for MNIST and SVHN or 70 clusters for CIFAR10.

6. Experiments

In this section we evaluate the different components that we have considered for clustering.

Table-A I-1 **Transformation-based Regularization**. We consider the information-based losses presented in section 4.1 and report results on the three datasets validation sets with different regularisation transformations.

Method	MNIST	CIFAR10	SVHN
$I(X,Y)$	42.6 ± 3.9	16.6 ± 1.0	15.4 ± 1.5
$I(X,Y)$ + $D_{\mathrm{KL}}(p(Y\|X), p(Y\|\mathrm{Geo}(X)))$	43.8 ± 47.8	14.5 ± 0.7	17.4 ± 0.8
$I(X,Y)$ + $D_{\mathrm{KL}}(p(Y\|X), p(Y\|\mathrm{VAT}(X)))$	**97.7** ± 0.3	18.4 ± 0.4	**17.4** ± 1.2
$I(X,Y)$ + $D_{\mathrm{KL}}(p(Y\|X), p(Y\|\mathrm{Mixup}(X)))$	34.4 ± 12.0	**20.1** ± 1.2	15.5 ± 1.3
$I(X,Y)$ + $D_{\mathrm{KL}}(p(Y\|X), p(Y\|\mathrm{Cuout}(X)))$	60.7 ± 9.0	18.1 ± 2.2	16.2 ± 1.5
$I(X,Y)$ + $D_{\mathrm{KL}}(p(Y\|X), p(Y\|\mathrm{Gauss}(X)))$	65.8 ± 12.0	17.2 ± 1.3	13.4 ± 0.4
$I(Y,\widetilde{Y})$	**97.9** ± 0.0	31.9 ± 1.1	28.0 ± 4.0
$I(Y,\widetilde{Y})$ + $D_{\mathrm{KL}}(p(Y\|X), p(Y\|\mathrm{VAT}(X)))$	97.6 ± 0.1	37.5 ± 2.9	34.3 ± 3.9
$I(Y,\widetilde{Y})$ + $D_{\mathrm{KL}}(p(Y\|X), p(Y\|\mathrm{Mixup}(X)))$	94.1 ± 6.7	**39.2** ± 0.9	27.0 ± 0.8
$I(Y,\widetilde{Y})$ + $D_{\mathrm{KL}}(p(Y\|X), p(Y\|\mathrm{Cutout}(X)))$	97.9 ± 1.3	33.9 ± 3.7	**39.6** ± 2.9
$I(Y,\widetilde{Y})$ + $D_{\mathrm{KL}}(p(Y\|X), p(Y\|\mathrm{Gauss}(X)))$	94.4 ± 2.5	29.7 ± 0.9	35.2 ± 3.3

First of all, in the next subsection we evaluate the clustering performance of the two mutual information losses. Next, we report the effect of adding a regularization term with different image transformations to the analyzed losses. Then, we consider the effect of using different transformations in the computation of the mutual information between an image and its transformed version. Finally, we show that the proposed clustering can be used to initialize the parameters of a network to improve image classification. We compare our best model with several well-known methods for disentangling representation on the task of linear supervised classification using the representation learned by the respective methods.

6.1 Mutual Information Loss

Table I-1 report results on the two ways of using mutual information for clustering as explained in section 4.1. $I(X,Y)$ considers the loss between input image and the output cluster. While this form of using the entropy is the most commonly used for clustering data (Bridle *et al.*, 1992; Krause *et al.*, 2010), we notice a large gap in performance compared to $I(Y,\widetilde{Y})$. This is probably because the latter already includes in the mutual information optimization a (geometrical) transformation, which brings more prior knowledge on the problem. In the next subsection we analyze the same two methods when explicitly adding an additional regularization term.

6.2 Transformation-based Regularization

In this section we consider the performance of the two information based losses when adding regularization based on the different transformations described in section 4.3. For each loss and each dataset we report in bold the transformation that leads to the best accuracy. For both losses $I(X,Y)$ and $I(Y,\widetilde{Y})$, adding a regularisation term based on D_{KL} seems to help.

For $I(Y,\widetilde{Y})$ in the table we consider $\widetilde{Y}$ as the marginal probabilities for samples with geometrical transformations.

Even though there is not a clear winner, a regularization based on VAT seems to lead to high performance on all datasets. For the following experiments we then use a model that maximises the mutual information between and example and its geometrical transformation and an additional regularization based on VAT.

In Fig. I-1 we visually compare the clustering performance of different mutual information losses with or without regularization. Each row, represent a cluster found in an unsupervised way. If the samples in each row belong to the same class, the clustering has managed to find a meaningful grouping strategy and the associated classification accuracy will be high. From visual inspection and similarly to the accuracies in table I-1, we observe that the clusters obtained with regularization have better similarity with semantic classes.

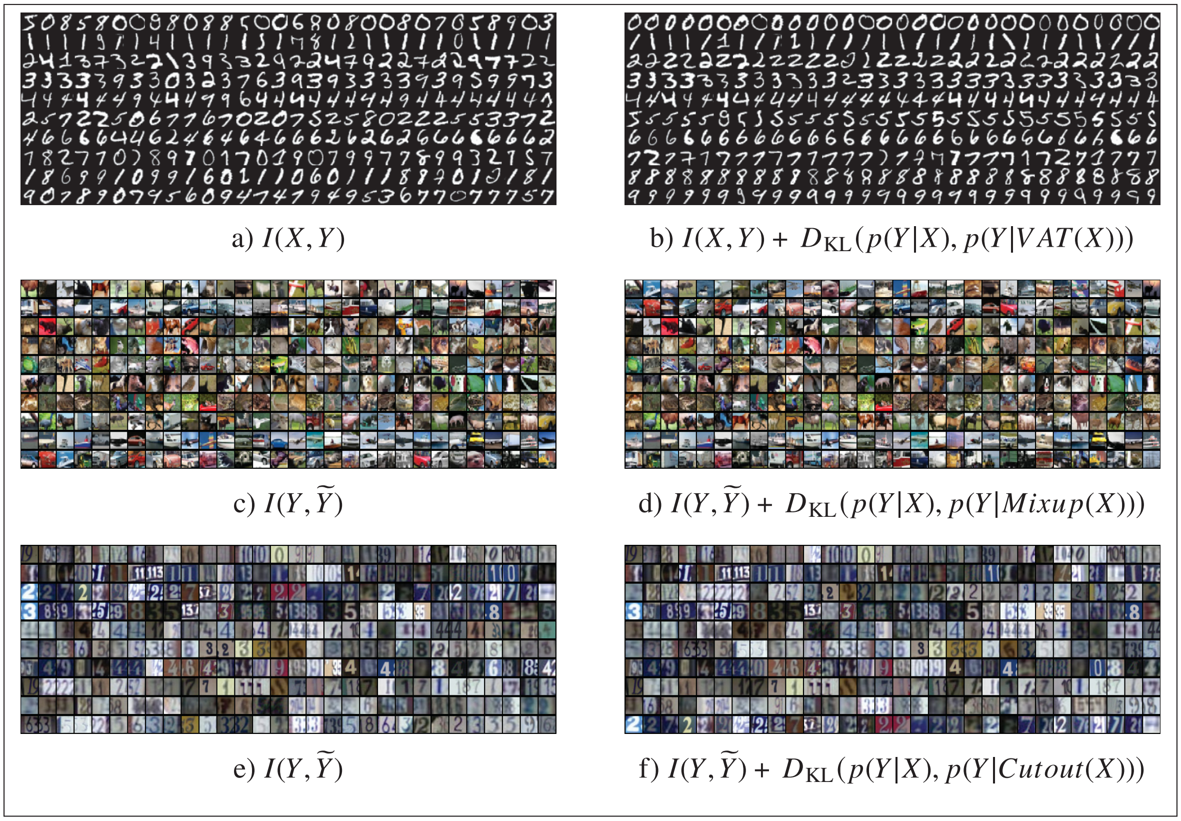
6.3 Transformations on the Mutual Information

In the previous section we have seen that maximizing the mutual information between and image and its transformed version $I(Y,\widetilde{Y})$ leads to better performance than maximizing the mutual information between input and output $I(X, Y)$ as in Hu *et al.* (2017b). Therefore here we focus on the first method and evaluate its performance when changing the transformations on $\widetilde{Y}$. In Table I-2 we see that when applying transformations directly on the mutual information loss, results are quite low, and the only adequate transformation seems to be the geometrical transformation originally used in (Ji *et al.*, 2019). This is because a transformation in the mutual information has a different effect than a transformation used for regularization. In regularization, we assume that the original image and the transformed image should contain the same class label. Thus, with KL divergence we impose that the output class probabilities of the image and the transformed should be similar. This is a way to make the loss smoother and therefore improve the optimization. In contrast, when using a transformation on the mutual information formulation does not require to maintain the same label between the original and the transformed image, it just maximizes their mutual information. Thus, in this setting stronger transformations seem to be required in order to obtain good results. This is verified by the fact that the same geometrical transformations, when used as regularization performed quite poorly (see Table I-1). This is because they are too strong to be used for regularization with KL divergence. For instance on MNIST, the strong crops of the image used in these geometrical transformation could change a 9 into a 0. Thus, for mutual information new and stronger transformations than those generally used for data augmentation are needed for improved results. We leave this as future work.

Among the used transformations, we noticed that VAT performed very poorly. This is because VAT is optimised to find adversarial samples for cross-entropy. Instead, in this case, we would like to generate and use samples that are adversarial for the mutual information loss. Thus, we change the adversarial generation in equ. A I-7, by using as a divergence the mutual information. In this case (IVAT) results are much more competitive.

Table-A I-2 **Transformations on the mutual information**. We consider the effect of changing the transformations used in the mutual information loss report results on the three datasets validation sets.

Method	MNIST	CIFAR10	SVHN
$I(Y, \widetilde{Y}_{Geo})$	**97.9** ± 0.0	**31.9** ± 1.1	**28.0** ± 4.0
$I(Y, \widetilde{Y}_{VAT})$	11.3 ± 2.2	17.6 ± 0.3	19.9 ± 0.2
$I(Y, \widetilde{Y}_{IVAT})$	70.1 ± 7.3	18.8 ± 2.5	19.1 ± 1.9
$I(Y, \widetilde{Y}_{Mixup})$	59.5 ± 5.6	20.3 ± 0.2	19.1 ± 1.7
$I(Y, \widetilde{Y}_{Cutout})$	33.5 ± 1.5	17.1 ± 1.3	13.3 ± 0.3
$I(Y, \widetilde{Y}_{Gauss})$	60.5 ± 4.4	17.1 ± 1.7	13.4 ± -0.1



a) $I(X, Y)$

b) $I(X, Y) + D_{\mathrm{KL}}(p(Y|X), p(Y|VAT(X)))$

c) $I(Y, \widetilde{Y})$

d) $I(Y, \widetilde{Y}) + D_{\mathrm{KL}}(p(Y|X), p(Y|Mixup(X)))$

e) $I(Y, \widetilde{Y})$

f) $I(Y, \widetilde{Y}) + D_{\mathrm{KL}}(p(Y|X), p(Y|Cutout(X)))$

Figure-A I-1 **$\mathbf{I(Y, \widetilde{Y})}$ with and without regularization.** We visually compare the obtained clusters for the three datasets with or without additional regularization. Each row represent a class in the dataset. Images have been randomly selected.

Table-A I-3 **Clustering for pre-training.** We evaluate how our best model can be used as initialization for supervised learning.

Training	DA	Accuracy
Scratch	no	72.9%
Our Pre-training	no	**80.6**%
Scratch	yes	81.6%
Our Pre-training	yes	**82.8**%

Table-A I-4 **Supervised Linear Classification**. To compare our model with self-supervised approaches, we extract a representation from the fully connected layer (FC), the penultimate convolutional layer (Conv) and the output (Y) of our clustering network and use it for supervised training on CIFAR10.

Method	FC	Conv	Y
VAE (Kingma & Welling, 2013)	42.1	53.8	39.6
AAE (Makhzani, Shlens, Jaitly, Goodfellow & Frey, 2015)	43.3	55.2	37.8
BiGAN (Donahue, Krähenbühl & Darrell, 2016)	38.4	56.4	44.9
DeepInfoMax (Hjelm *et al.*, 2018)	54.1	63.3	49.6
Ours	**75.4**	**78.9**	**64.7**

6.4 Comparison with IMSAT and IIC

In table I-5 we compare our best configuration with the results and setting presented in the IMSAT and IIC original papers on MNIST, CIFAR10 and SVHN. The comparison with IMSAT is reported in the first three rows of the table. Our re-implementation of IMSAT correspond to $I(X, Y) + D_{\mathrm{KL}}(p(Y|X), p(Y|\mathrm{VAT}(X)))$ as presented in table I-1. We can see that for MNIST our values are slightly lower than the original paper. This can due to the fact that in the original implementation of IMSAT, for estimating the correct amount of noise ϵ to use, they used an adaptive formulation based on the distances between samples in pixel space. This approach works only when distances in the input space are meaningful (thus not in difficult datasets). For datasets other than MNIST, in the original paper (Hu *et al.*, 2017b) clustering is not performed directly on images, but in more meaningful representations (ResNet50 feature maps pretrained on ImageNet). Instead, as our aim is to compare in fair way different methods, we use raw images on all experiments. Thus, we consider ϵ a hyper-parameter parameter. This explains why our performance on CIFAR and SVHN are lower than in the original paper. To further verify that our implementation is competitive we also run an experiment of our IMSAT implementation with input features extracted from ResNet50 pretrained on ImageNet as in the original IMSAT paper. For this experiment on CIFAR10 we obtained an accuracy of 75.5, which is significantly higher than the 45.6 of the original paper. We also compare our re-implementation of IMSAT with our best model. To make the comparison fair, we use our best model with 1 head and 1 subhead (1H1S) as in the original IMSAT paper. For MNIST our model obtain an accuracy that is slightly inferior to IMSAT. However, for more difficult datasets (CIFAR10 and SVHN) our model clearly outperform our IMSAT baseline.

For IIC we compare the performance of our best model with the values reported in the original paper. In this case, for fair setting we use a model with 2 heads and 5 over-clustering (70 clusters) subheads (2H5S) as in the original IIC paper. In this case our model obtained an accuracy that is slightly lower than the original paper on MNIST, but significantly better on CIFAR10.

Table-A I-5 **Comparison with IMSAT and IIC.** * means that the clustering was effectuated on higher level features from which is easier to cluster, therefore a direct comparison of the accuracies is not fair.

Method	MNIST	CIFAR10	SVHN
IMSAT (Hu *et al.*, 2017b)	98.4 ± 0.4	45.6 ± 0.8*	57.3 ± 3.9*
IMSAT (our impl.)	97.7 ± 0.3	18.4 ± 0.4	17.1 ± 2.0
Our Model (1H1S)	97.6 ± 0.1	37.5 ± 2.9	34.3 ± 3.9
IIC (Ji *et al.*, 2019)	98.4 ± 0.7	57.6 ± 5.0	-
Our Model (2H5S)	94.9 ± 2.1	62.5 ± 0.3	38.6 ± 0.8

6.5 Clustering for pre-training

In table I-3, we consider the effect of using the network trained for clustering as pre-training unsupervised initialization for a fully-supervised training on the same dataset (CIFAR10). When we train without any data augmentation, the initialization based on clustering gives an important boost in performance, going from an accuracy of 72.9% for a network trained from scratch to 80.6% for a network whose weights are initialized with our best model. When the network is trained with data augmentation, the gap between the two initialization is reduced to 1.2%. Thus, there is still a gain without introducing any additional data or annotations, but more limited. This is probably due to the fact that most of the useful information brought in by the clustering pre-trained model is about image transformations. Thus, when adding data augmentation that information is included directly in the training and the initialization becomes less useful.

6.6 Disentangled Representations

As we perform clustering, the final results are already groups of samples that are visually similar. If the clustering works well, these groups should represent classes and they can be directly used for classification. Thus, in our experiments, it is not necessary to perform any additional supervised learning for evaluation. In contrast, methods based on self-supervision or that learn disentangled representations will normally need a final evaluation step in which supervision is used to learn the final classifier. An evaluation often used for these methods consists on training the learned representation in a fully-supervised way but with a linear classifier so that its capacity is limited. In order to compare with methods based on self-supervision, we use our learned representation to train a linear classifier. We argue that, as clustering is very similar to the final classification task, results of our method will be better than other approaches. Table I-4 compares DeepInfoMax (Hjelm *et al.*, 2018), a recent method for disentangled representations based on mutual information, with our clustering approach. We also report results of other popular methods from (Hjelm *et al.*, 2018). We report the accuracy obtained by a linear classifier trained on the fully-connected layer (FC), penultimate convolutional layer (Conv) and output (Y), 10 values in our case and 64 for DeepInfoMax. As expected, the gap in performance is quite high: in the order of 20 points for FC, 10 for convolutional and 5 for the output Y. The reduced gap in the output is explained by the low dimensionality of the final output (10) for our model.

7. Conclusions

In this paper, we have presented a generalization of two very popular information-based clustering approaches. We consider a clustering model based on a combination of a mutual information based loss and a regularization based on transformations. From our empirical evaluation we conclude that adding a regularization based on image transformations is in most of the cases beneficial. Our best configuration is then a clustering that maximizes the mutual information between a samples and its geometrical transformation, regularised by a KL divergence between a sample and its adversarial transformation. This configuration seems to outperform previous work on clustering as well as on disentangling representations.

APPENDIX II

DIVERSIFIED MULTI-PROTOTYPE REPRESENTATION FOR SEMI-SUPERVISED SEGMENTATION

Jizong Peng[1] , Marco Pedersoli[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] Department of Automated Production, École de Technologie Supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3

Article accepted to *Medical Imaging Meets NeuRIPS 2021 Workshop* [1] in Nov. 2021

1. Presentation

This appendix presents “Diversified Multi-prototype Representation for Semi-supervised Segmentation", a paper accepted to *Medical Imaging Meets NeuRIPS 2021* workshop. In this work, we consider semi-supervised segmentation as a dense prediction problem based on prototype vector correlation, and proposes a simple way to represent each segmentation class with multiple prototypes. To avoid degenerate solutions, two regularization strategies are applied on unlabeled images. The first one leverages mutual information maximization to ensure that all prototype vectors are considered by the network. The second explicitly enforces prototypes to be orthogonal by minimizing their cosine distance. Experimental results on two benchmark medical segmentation datasets reveal our method’s effectiveness in improving segmentation performance when few annotated images are available.

2. Introduction

Supervised learning approaches based on convolutional neural networks (CNNs) have achieved outstanding performance in a wide range of segmentation tasks. However, these approaches often

[1] https://sites.google.com/view/med-neurips-2021/abstracts?authuser=0

require a large amount of labeled images, difficult to obtain for medical imaging applications. Semi-supervised learning (SSL) is commonly used to reduce the need for fully-annotated data by exploiting unlabeled ones. In recent years, a broad range of semi-supervised methods have been proposed for medical image segmentation, including approaches based on transformation consistency (Perone & Cohen-Adad, 2018; Bortsova *et al.*, 2019; Luo *et al.*, 2020), co-training of multiple models (Peng *et al.*, 2020a; Wang *et al.*, 2021; Zhou *et al.*, 2018a) and adversarial learning (Decourt & Duong, 2020; Li *et al.*, 2020b). While different from image classification, segmentation can be viewed as a dense prediction problem where classification is performed at the pixel level. Based on this idea, recent works have used metric learning (Hoffer & Ailon, 2015; Chaitanya *et al.*, 2020) to learn a robust representation for this task. These methods were shown to be effective for cardiac (Peng *et al.*, 2021b; Chaitanya *et al.*, 2020; Xiang, Li, Wang, Xia & Zhang, 2021) and prostate (He *et al.*, 2021; Chaitanya *et al.*, 2020) segmentation. The core principle of these methods is to minimize a distance metric for similar pixels (*positive* pairs) while minimizing it for dissimilar ones (*negative* pairs). These pairs can be defined by image region (Xiang *et al.*, 2021; Chaitanya *et al.*, 2020; Peng *et al.*, 2021b) or segmentation boundary (He *et al.*, 2021).

In this work, we consider segmentation as the problem of learning a fixed number of prototype vectors, modeled as 1×1 convolution kernels in the last convolution layer. During training, prototype vectors of a given class are updated so that they have high correlation to feature vectors located in regions corresponding to that class. We argue that widely-used segmentation networks such as U-Net (Ronneberger *et al.*, 2015) are designed to have only one prototype vector per class, and that this may be insufficient to capture the full variability of a class. Instead, we assign multiple prototype vectors to the same class by increasing the number of output classes. This simple strategy leads to an over-segmented prediction that can then be converted to a normal segmentation by learning the right prototype-to-class correspondence. To avoid trivial solution for the over-segmented prediction, we introduce two regularization losses based on mutual information maximization and prototype vector orthogonality. The former one ensures that prototypes are used in a fair manner, while the other imposes them to be uncorrelated. These losses increase the usefulness and diversity of prototypes, which leads to a more robust prediction. We note that this differs from the Prototypical Network approach (Snell, Swersky & Zemel, 2017) for few-shot learning, where prototype vectors are obtained by averaging the representations of examples in the same class, similar to K-Means (Ren *et al.*, 2018). We validate our proposed method on two benchmark medical image datasets, evaluating the binary and multi-class segmentation of different organs. Results show our method to outperform various baselines and recent approaches for semi-supervised segmentation, when few labeled images are provided.

3. Method

In semi-supervised segmentation, we are given a small set labeled examples $\mathcal{D}_\ell = \{(x_\ell,_\ell)\}$, each composed of an image $x \in \mathbb{R}^{|\Omega|}$ and segmentation ground-truth $_\ell \in \{0,1\}^{|\Omega|\times|C|}$, and a larger set of unlabeled images $\mathcal{D}_u = \{x_u\}$. Here, $\Omega = \{1,\ldots,W\times H\}$ is the set of image pixels indexes and C the number of segmentation classes. We seek a segmentation function s with parameters θ that takes as input an image x and returns for each pixel i and class label k a probability $s_{ik}(x)$. A CNN-based network is typically used for s. Such architecture decomposes the segmentation function as $s = (g \circ \phi)$, where $\phi(x) \in \mathbb{R}^{|\Omega|\times N}$ is a feature map obtained obtained from CNN layers and g is a function projecting the features to class probabilities, typically implemented using a 1×1 convolution followed by a softmax. The final output can thus be expressed as $s_{ik}(x) = \text{softmax}(\mathbf{w}_{\mathrm{k}}^\top \text{Œ}_{\mathrm{i}}(\mathrm{x}))$ where $\mathbf{w}_k$ is the k-th column of a kernel matrix $\mathbf{W} \in \mathbb{R}^{N\times C}$. As in most few shot learning approaches (Snell *et al.*, 2017), we can consider $\mathbf{w}_k$ as a prototype vector for class k. The probability of pixel i to be mapped to class k is then proportional to the correlation between its feature vector $\phi_i(x)$ and prototype $\mathbf{w}_k$.

A problem with this simple model is its strong assumption that the variability of a class can be fully captured with a single prototype. We argue that this limited class representation is suboptimal, especially when few annotated images are available for training. A straightforward way to alleviate this problem, which has not been well explored in the medical imaging community, is to represent each class with multiple prototype vectors. Toward this goal, we augment the number of output classes from C to C', with $C' > C$, which leads to an over-segmentation of the input image. Denote as $s'(x) \in [0,1]^{|\Omega| \times C'}$ this new output, one can then recover the final segmentation by finding a mapping : $\{1,\ldots,C'\} \rightarrow \{1,\ldots,C\}$ from the over-segmentation classes to the real ones. While this mapping can be learned, in our method, we assume that each class is represented by the same number of prototypes and define as a fixed $C' \times C$ matrix $\mathbf{M}$ such that $m_{jk} = 1$ if over-segmentation class j is mapped to real class k, else $m_{jk} = 0$. The probability for pixel i and class k is then obtained as $s_{ik}(x) = \sum_{j=1}^{C'} m_{jk} f'_{ij}(x)$. We leverage annotated images with a supervised loss measuring the cross-entropy between the ground-truth labels and the real class probabilities:

$$\mathcal{L}_{\mathrm{sup}}(\theta; \mathcal{D}_\ell) = \frac{1}{|\mathcal{D}_\ell|} \sum_{(x,) \in \mathcal{D}_\ell} \ell_{\mathrm{CE}}(, f(x)), \text{ where } \ell_{\mathrm{CE}}(, \mathbf{p}) = -\frac{1}{|\Omega|} \sum_{i=1}^{|\Omega|} \sum_{k=1}^{C} y_{ik} \log(p_{ik}). \quad \text{(A II-1)}$$

However, optimizing this supervised loss does not guarantee that the prototypes of over-segmentation classes model distinct and useful properties of their underlying real class. To ensure this, we add two regularization losses on unlabeled examples, based on mutual information maximization and prototype orthogonality.

Mutual Information based regularization Since the number of over-segmentation prototypes exceeds the number of real classes, it can happen that some of these prototypes are ignored during optimization and, thus, have a low output probability. To ensure that all prototypes are used and that their marginal distribution is balanced, we maximize the mutual information $I(X;Y')$ between a random variable X modeling an image and a random variable Y' encoding its over-segmentation class label. Using the fact that $I(X;Y') = \mathcal{H}(Y') - \mathcal{H}(Y'|X)$, where

$\mathcal{H}(\mathbf{p}) = -\frac{1}{|\Omega|}\sum_i \sum_k p_k \log(p_{ik})$ is the Shannon entropy (averaged over pixels), we define the first regularization loss as

$$\mathcal{L}_{\mathrm{MI}}(\theta; \mathcal{D}_u) \;=\; -\mathcal{H}\Bigg(\frac{1}{|\mathcal{D}_u|}\sum_{x \in \mathcal{D}_u} s'(x)\Bigg) \;+\; \frac{1}{|\mathcal{D}_u|}\sum_{x \in \mathcal{D}_u} \mathcal{H}(s'(x)). \tag{A II-2}$$

As described in Hu *et al.* (2017a), minimizing the first term of $\mathcal{L}_{\mathrm{MI}}$ increases the entropy of the marginal distribution, defined as the average of individual predictions, making it more uniform. On the other hand, the second term encourages the network to have confident predictions for the over-segmentation classes.

Orthogonality regularization To learn prototype vectors representing distinct characteristics, we also impose them to be uncorrelated. This is achieved with our prototype orthogonality loss:

$$\mathcal{L}_{\mathrm{orth}}(\theta) \;=\; \|\mathbf{W}\mathbf{W}^\top - \mathbf{I}_{C'}\|_F^2 \tag{A II-3}$$

where $\mathbf{I}$ is the $C' \times C'$ identity matrix.

Complete objective The total loss, including the supervised loss and the two unsupervised regularization losses, is finally defined as

$$\mathcal{L}_{\mathrm{total}} \;=\; \mathcal{L}_{\mathrm{sup}}(\theta; \mathcal{D}_l) \;+\; \lambda_1 \mathcal{L}_{\mathrm{MI}}(\theta; \mathcal{D}_u) \;+\; \lambda_2 \mathcal{L}_{\mathrm{orth}}(\theta) \tag{A II-4}$$

where $\lambda_1, \lambda_2 \geq 0$ are coefficients balancing the different loss terms.

4. Experiments

Dataset and evaluation metric We evaluate our proposed method and its variants on two benchmark datasets for medical image segmentation: Promise12 and ACDC. Promise12 consists of 50 prostate MRI scans from different patients, which are split randomly in a training, Table-A II-1 Mean 3D DSC over three independent runs of tested methods.

Method	**MP**	**MI**	**Orth**	**PROMISE12**		**ACDC**		
				3 scans	**5 scans**	**1 scan**	**2 scans**	**4 scans**
Full Sup.				88.06		89.70		
Baseline				52.33	65.17	60.77	74.50	78.91
Variant-1	✓			53.20	70.95	65.25	75.95	80.56
Variant-2		✓		71.52	79.50	69.91	80.75	81.44
Variant-3			✓	52.55	66.55	64.53	74.15	80.74
Variant-4	✓	✓		**73.88**	78.30	72.00	80.68	83.01
Variant-5	✓		✓	55.01	70.95	66.58	77.17	78.97
Ours	✓	✓	✓	73.38	**79.69**	**72.44**	**82.11**	**83.09**
Entropy Min. Vu *et al.* (2019)				53.37	69.74	67.53	76.08	80.59
Pseudo Label Li, Ko & Choi (2019)				53.50	67.40	65.74	75.33	79.59
Mean Teacher Perone & Cohen-Adad (2018)				65.73	78.79	81.61	84.00	**84.72**
MT + Ours	✓	✓	✓	**75.30**	**80.26**	**82.10**	**85.70**	84.53

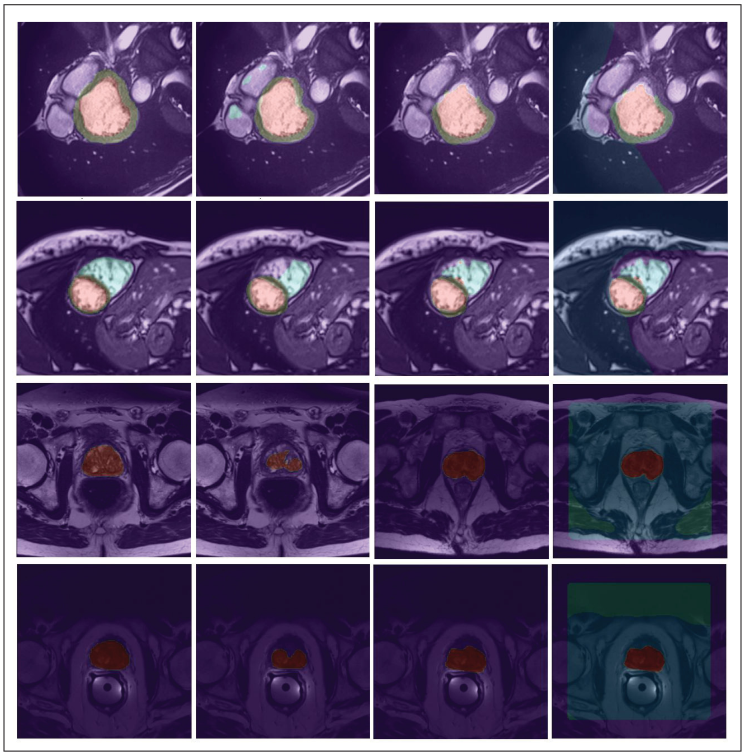


Figure-A II-1 Visual results. From left to right: ground truth, baseline, ours, and over-segmented prediction.

validation and test sets containing 40, 3, and 7 scans, respectively. ACDC contains 200 cardiac cine-MRI scans and corresponding ground-truth segmentation masks for three classes: left ventricle (LV), right ventricle (RV) and myocardium of LV. We randomly selected 174, 9, and 17 scans of this dataset for training, validation and testing. The same prepossessing protocol and data augmentation were used as in Peng *et al.* (2021a). For semi-supervised learning, we randomly chose a few scan from their training set as labeled set and considered others as unlabeled. We report the mean 3D Dice score (DSC) on the test set over 3 runs.

Implementation details We choose the 2D UNet (Ronneberger *et al.*, 2015) as our main architecture and optimize it using stochastic gradient descent (SGD) with the RAdam optimizer (Liu *et al.*, 2019). For all experiments, we used the public code base [2] from Peng *et al.* (2021a), keeping the same optimization strategy. Using a grid search on the validation set, we selected the following hyper-parameters: 3 prototype vectors per class, $\lambda_1 = 0.01$ and $\lambda_2 = 0.5$.

Results The upper part of Table II-1 reports the performances of baselines and ablation variants of our method. *Full Sup.* is an upper bound where all available training data are used as labeled examples. Likewise, *Baseline* serves as our lower bound where only a few labeled data are used (see Table) and unlabeled images are ignored. We tested six variants of our method using multiple prototypes or a single one (MP), the mutual information loss (MI), and the prototype orthogonality loss (Orth). As can be seen, using MP, MI, and Orth together leads to the greatest improvement over the Baseline in all but one cases, boosting DSC by 21.1% for Prostate12 with 3 labeled scans, and by 11.7% for ACDC with one labeled scan. Visual results are given in Fig. II-1, showing examples of the predicted over-segmentation and its benefit on the final segmentation. We then compare our method to three popular semi-supervised segmentation approaches, (Vu *et al.*, 2019), pseudo-label (Li *et al.*, 2019), and Mean Teacher (MT) (Perone & Cohen-Adad, 2018). Results show our method to outperform Entropy minimization and pseudo-label for both datasets, while MT is better for ACDC. Based on this observation, we boost our method by adding a temporal ensembling strategy on top as in

[2] https://github.com/jizongFox/MI-based-Regularized-Semi-supervised-Segmentation

Mean Teacher. Results at the bottom of Tab. II-1 show a significant improvement over standard Mean Teacher when using very few labeled images.

5. Social impact

The proposed method can have a practical impact on clinical applications by reducing the need for fully-annotated data. As shown in our experiments, it improves the accuracy of medical image segmentation when very few labeled images are available. This could help reduce the workload of radiologists, thereby reducing costs, and provide clinicians with better information for diagnosis. While our empirical evaluation has shown excellent results with very limited data, using fewer annotated images also increases chances of over-fitting outliers in the data (e.g., poor annotations) and produce erroneous or misleading results. A further study on the reliability of medical image segmentation with reduced images is therefore recommended.

APPENDIX III

SUPPLEMENTARY MATERIAL FOR «SELF-PACED CONTRASTIVE LEARNING FOR SEMI-SUPERVISED MEDICAL IMAGE SEGMENTATION WITH META-LABELS»

1. Presentation

This appendix offers detailed explanation of the experimental setup, dataset description and complementary experimental results for “Self-Paced Contrastive Learning for Semi-supervised Medical Image Segmentation with Meta-labels", which was accepted at *NeuRIPS 2021* conference and presented in chapter 5.

2. Conceptual diagram and algorithm flow of proposed method

We illustrate the principle of our proposed self-paced contrastive learning mechanism in Fig. III-1, where slice position is used as the meta-label. Our loss is computed on top of the conventional contrastive loss, while considering the self-paced coefficient w_{ij} given a batch of unlabeled images. The self-paced coefficient w_{ij}, which measures thee reliability of a positive pair (i, j), is calculated as in Equ. (6). We also include in **Algorithm** III-1 the algorithm flow of our method used with arbitrary meta-labels.

3. Dataset description

- **ACDC dataset** This publicly-available dataset (Bernard *et al.*, 2018) contains 200 short-axis cine-MRI scans obtained from 100 patients. Scans were acquired using 1.5 and 3 T systems with an in-plane resolution from 0.70 × 0.70 mm to 1.92 × 1.92 mm and a through-plane resolution from 5 mm to 10 mm. Volumetric images were obtained for end-diastolic (ED) (100 scans) and end-systolic (ES) (100 scans) phases of the cardiac cycle. Ground-truth segmentation masks are provided for the following four regions of interest: left ventricle

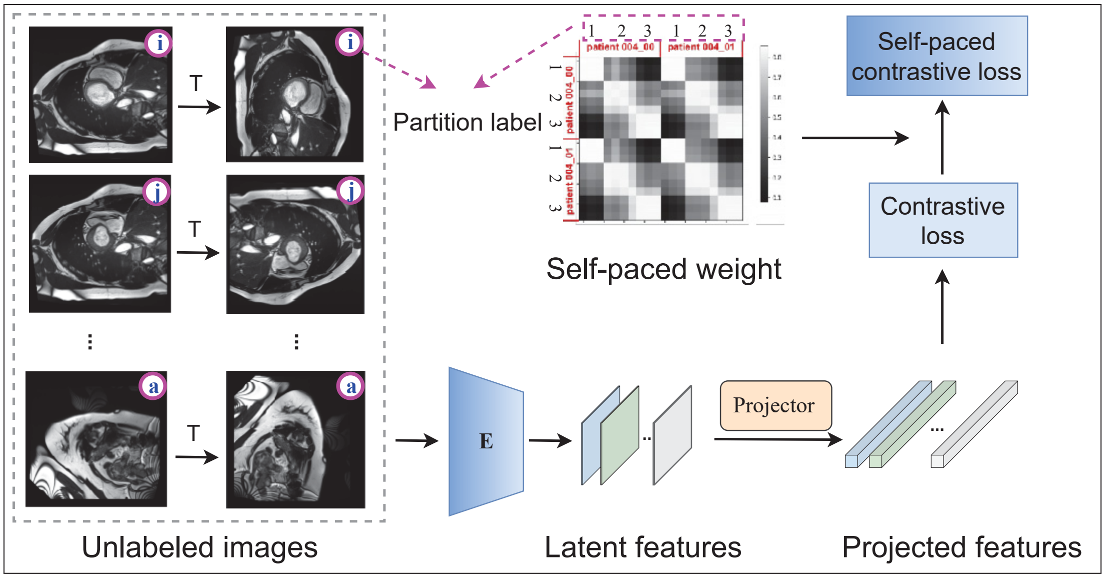


Figure-A III-1 Conceptual diagram for our proposed self-paced contrastive learning method with meta-labels.

endocardium (LV), left ventricle myocardium (Myo), right ventricle endocardium (RV), and background. Due to a highly-variable resolution, we slice 3D scans through-plane into 2D images, and re-sample these 2D images to a fixed resolution of 1.0 × 1.0 mm. For each scan, intensities were normalized using the 1% and 99% percentile of the scan's intensity histogram before performing slicing. Normalized 2D images are then cropped to a size of 384 × 384. For our experiments, we used a random split of 175 scans as our training set, from which we randomly select 1, 2 or 4 scans as our labeled data, and considered others as unlabeled images[1]. We then divided the remaining 25 scans into a validation set consisting of 8 scans and a test set with 17 scans. Both the validation and test sets were set aside during model optimization. We employed a diverse set of data augmentations during training, for both labeled and unlabeled images, which include random crops of 224 × 224 pixels, random flips, random rotations within $[-45, 45]$ degrees, and color jitters.

We leveraged the rich meta information available in ACDC to obtain three types of meta-labels.

[1] Once selected, we fixed these splits across all experiments.

Algorithm-A III-1 Self-paced contrastive learning in pretraining stage.

```
Input: Unlabeled dataset $\mathcal{U}$ and their respective meta-label set $D_{\text{meta}}$; Encoder of the
       segmentation network $E(\cdot)$; Temperature $\tau$; Learning pace $\gamma$ scheduler ;
Output: Pre-trained model parameters $\{\theta\}$ for $E(\cdot)$;
Initialize network parameters $\theta$;
Initialize hyper-parameters: learning pace: $\gamma \leftarrow \gamma_0$ = Scheduler (0) ;
for epoch = 1, ..., $n_{\text{epochs}}$ do
    for n = 1, ..., $n_{\text{iter}}$ do
        Sample unlabeled training batch $\{\mathcal{U}_n\}$;
        For all $x_u \in \mathcal{U}_n$, do random transformation and get $x_u^T$;
        Compute $z$ using non-linear project $g(\cdot)$ for the features $E(x_u)$;
        Compute the sample-wise contrastive loss using Eq. (2):
        $\ell_{ij} = -\log \frac{\exp\left(\mathbf{z}_i^{\top}\mathbf{z}_j/\tau\right)}{\sum_{a\in\mathcal{A}(i)}\exp\left(\mathbf{z}_i^{\top}\mathbf{z}_a/\tau\right)}$;
        Compute self-paced importance weight $\omega_{ij}$ using Eq. (6):
        $w_{ij}^{*} = \underset{w_{ij}\in[0,1]}{\arg\min}\; w_{ij}\,\ell_{ij} + R_{\gamma}(w_{ij})$;
        Compute self-paced contrastive loss using Eq. (3):
        $\mathcal{L}\textsc{sp}\text{-con}^k = \frac{1}{2N}\sum_{i=1}^{2N}\frac{1}{|P^k(i)|}\sum_{j\in P^k(i)} w_{ij}\,\ell_{ij} + R_{\gamma}(w_{ij})$;
        According to Eq. (14), do a batch gradient descent step on the model's parameters $\theta$;
        Update the model's parameters $\theta$;
    end for
    Adjust the SGD learning rate;
    Update learning pace according to the scheduler: $\gamma \leftarrow$ Scheduler (epoch);
end for
return $\{\theta\}$ ;
```

Following Chaitanya *et al.* (2020), we first considered slice position and patient identity as meta-labels in our experiments, referring respectively as $\mathcal{L}\text{con}/\textsc{sp}^1$ and $\mathcal{L}\text{con}/\textsc{sp}^2$ the standard contrastive loss and self-paced contrastive loss based on these meta-labels. As Chaitanya *et al.* (2020), we defined the position of a 2D image in a volume based on the partition of this volume into $S=3$ equal-sized groups of consecutive slices. Additionally, we used the cardiac phase of the scan (i.e., ED or ES) as third meta-label, and write as $\mathcal{L}\text{con}/\textsc{sp}^3$ the contrastive losses using this meta-label. We detail the metal labels in the next section.

- **PROMISE12 dataset** The second dataset used to evaluate our method focuses on prostate

MRI segmentation (Litjens *et al.*, 2014) . It comprises multi-centric transversal T2-weighted MR images from 50 subjects, acquired with scanners from multiple vendors and different scanning protocols. Image resolution ranges from $15 \times 256 \times 256$ to $54 \times 512 \times 512$ voxels with a spacing between $2 \times 0.27 \times 0.27$ and $4 \times 0.75 \times 0.75$ mm^3. We sliced scans into 2D images along the short-axis and resized these images to a size of 256×256 pixels. Intensities were once again normalized based on the 1% and 99% percentiles of the intensity histogram for each scan before the slicing operation. We randomly selected 40 scans as our training set, and used 3, 5 or 7 scans from this set as our labeled data. We considered a validation set with 4 scans and a test set with 6 scans. For data augmentation, we utilized the same set of transformations as with the ACDC dataset, except that we limit random rotations to $[-10,10]$ degrees. Similar to Chaitanya *et al.* (2020), we adopted slice position ($S = 5$ partitions) and patient identity as meta-labels for this dataset.

- **MMWHS dataset** The third dataset considered in our evaluation, the Multi-Modality Whole Heart Segmentation (MMWHS) dataset (Zhuang & Shen, 2016), consists of high resolution CT images from 20 subjects. Four segmentation classes were used in our experiments: left ventricle myocardium (LVM), left atrium blood cavity (LAC), left ventricle blood cavity (LVC) and ascending aorta (AA). Following a similar protocol as with the ACDC dataset, volumetric images were first normalized based on their intensity histogram, then sliced along the short-axis, and finally resized to a resolution of 256×256 pixels. We randomly selected 10 scans as our training set, from which 1 or 2 were used as our labeled data and the others as unlabeled data. Validation and test sets contain 4 and 6 scans, respectively. The same set of transformation as with the ACDC dataset was used to augment images on the fly during training. Once again, we adopted slice position ($S = 7$ partitions) and patient identity as meta-labels for this dataset.
- **Hippocampus dataset** The fourth dataset, as a sub-track of Medical Segmentation Decathlon (Antonelli *et al.*, 2021), aims to segment hippocampus from 260 T1-sequence MRI images

acquired from both healthy adults and adults with a non-affective psychotic disorder. As before, volumetric images were normalized according to their histogram and sliced to 2D images along the short axis with a spatial size of 96 × 96 pixels. We randomly split the images into training, validation and test set, consisting of 223, 12 and 25 scans, respectively. To perform semi-supervised segmentation, we then chose 1, 2, or 4 scans from these training examples as the labeled ones, while keeping others as unlabeled. We used the same data augmentation as the ACDC dataset and considered slice position (S = 3 partitions) as the meta-label.

- **Spleen dataset** The last dataset (Antonelli *et al.*, 2021) consists of patients undergoing chemotherapy treatment for liver metastases. A total of 41 portal venous phase CT scans were included in the dataset with acquisition and reconstruction parameters described in Antonelli *et al.* (2021). Similar to the previous datasets, 2D slices were obtained by slicing the high-resolution CT volumes along the axial plane. Intensities in each slice were clipped to a range of $[-100, 400]$ and resulting images resized to a resolution of 512 × 512 pixels. We randomly split the dataset into training, validation and test sets, comprising CT scans of 35, 2, and 5 patients respectively. To evaluate algorithms in a semi-supervised setting, we then randomly chose 2 or 4 scans from the training set as labeled examples and considered remaining images as unlabeled. We again applied the same set of data augmentations as for the ACDC dataset and employed slice position (S = 5 partitions) as the meta-label for this dataset.

4. Meta information visualization

In Fig. III-2-III-4, we visualize examples of images from the three first datasets, corresponding to different patients (ACDC, PROMISE12 and MMWHS), slice partition (ACDC, PROMISE12 and MMWHS), and phase of the cardiac cycle (ACDC only). Although a given slice partition exhibits a high-level structural similarity across different volumes, we also observe important variability in corresponding images, due to differences in acquisition conditions, individual anatomy of subjects and imperfect image registration. Thus, without human interaction, this meta-label can lead to the learning of noisy representations. The second meta-label, patient identity, reflects global differences between scans. These differences are particularly notable for the PROMISE12 and MMWHS datasets, where the structural shape and the image contrast differ significantly across patients. We experimentally show that using this global meta-label by itself may improve segmentation performance, however it is more useful when combined with other meta-labels like slice partition. Our third meta-label for the ACDC dataset is the cardiac cycle phase (i.e., ES or ED). A single cycle of cardiac activity can be divided into two basic phases, the diastole where the ventricles are relaxed (not contracting), and the systole where the left and right ventricles contract and eject blood into the aorta and pulmonary artery, respectively. The first two rows of Fig. III-2 show images corresponding to these two phases for the same patient. One can see that the size of the left and right ventricles changes considerably, and is much smaller at the end of the systole phase (ES). As shown in our experiments, incorporating information on the cardiac phase into the network's encoder generally helps improve segmentation quality.

4.1 Evaluation metric

We used the commonly-adopted Dice similarity coefficient (DSC) metric to evaluate segmentation quality of the tested methods. DSC measures the overlap between the predicted labels (S) and the corresponding ground truth labels (G):

$$\mathrm{DSC}(S,G) = \frac{2\,|S \cap G|}{|S|+|G|} \tag{A III-1}$$

DSC values range from 0 to 1, a higher value corresponding to a better segmentation. In all experiments, we reconstruct the 3D segmentation for each scan by aggregating the predictions made on 2D slice and report their 3D DSC metric on the test set.

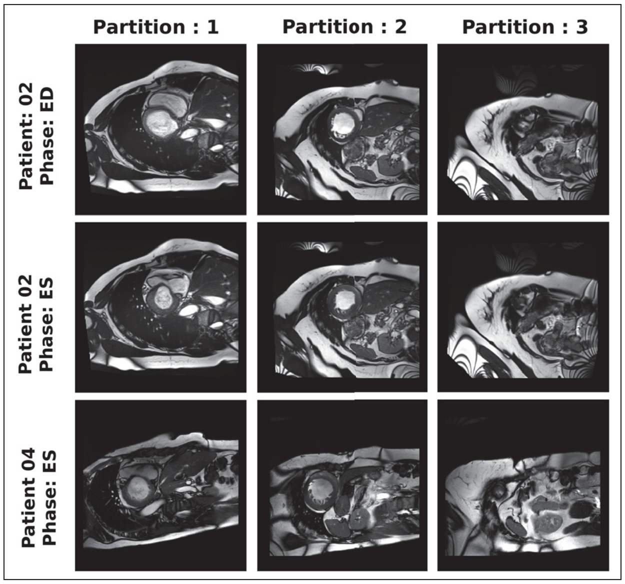


Figure-A III-2 Examples of ACDC images with meta-labels. Volumetric images are sliced thought the short-axis and split into $S = 3$ partitions. One can see that different slices in the same partition can share similar structure across volumes. We also consider the patient identity and the cardiac cycle phase as global meta-labels to guide the model optimization.

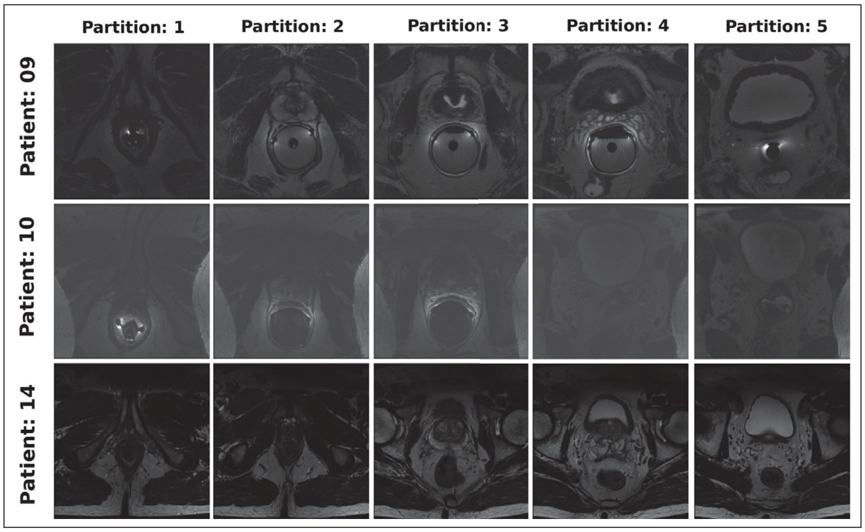


Figure-A III-3 Examples of PROMISE12 images with their meta-labels. Slice partition and patient identity are used as the two meta-labels. We fixed the number of partitions to $S = 5$. One can see a smooth transition on anatomical structures from a partition to the next. Note that this dataset exhibits high appearance variability across different patients.

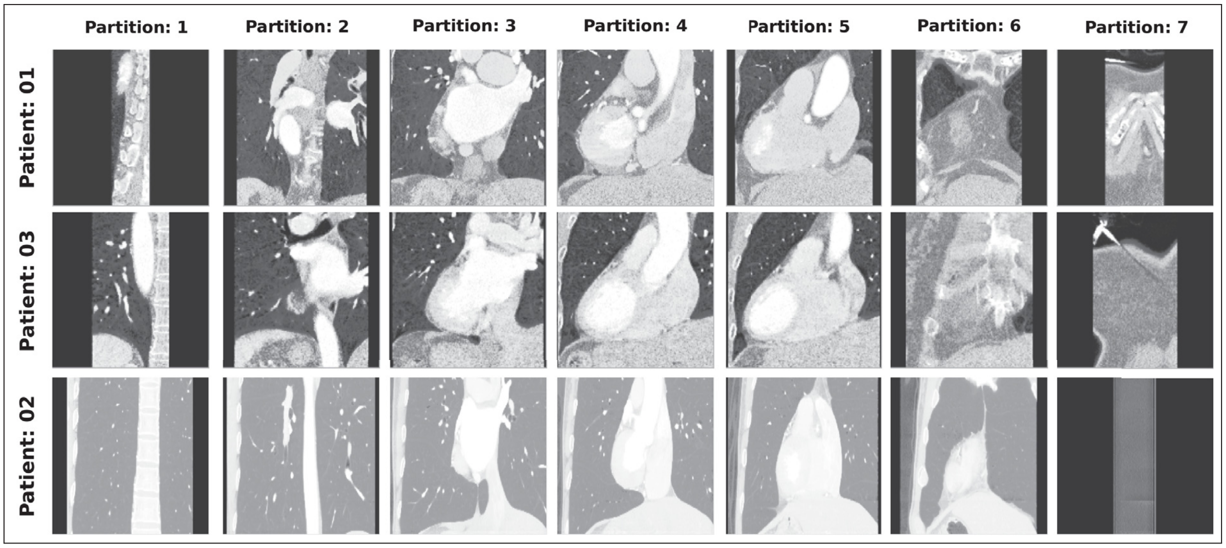


Figure-A III-4 MMWHS images with their meta-labels. Slice partition and patient identity are also used as the two meta-labels. We fixed the number of partitions to $S=7$. Smooth transition on anatomical structures can be observed from a partition to the next. This dataset also exhibits high appearance variability across different patients.

4.2 Experimental details

We assessed the performance of our self-paced contrastive learning approach when used in two different stages: **pre-training** and **semi-supervised learning**. The pre-training stage consists in optimizing the encoder of a segmentation network on all available images via the contrastive loss. An additional fine-tune step is usually appended to this setting, which trains the whole network on a few labeled scans. In contrast, semi-supervised learning trains the network jointly with both labeled and unlabeled images. In both cases, we applied a learning rate warm-up strategy to increase the initial learning rate lr_{int} by a factor of N in the first 10 epochs, and then decrease it with a cosine scheduler for the following `max_epoch` – 10 epochs, as shown in Fig. III-5.

For pre-training stage, lr_{int} was set to 5×10^{-5}, N as 400 and `max_epoch` as 80 for the PROMISE12, MMWHS, Hippocampus and Spleen datasets, whereas we used `max_epoch` = 120 for ACDC. For the semi-supervised training stage or the fine-tuning procedure, we set lr_{int} to

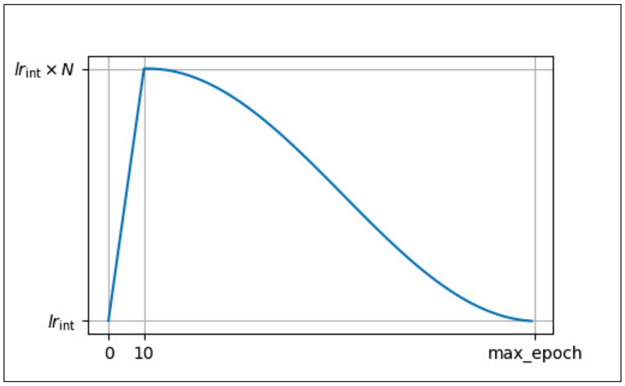


Figure-A III-5 Learning rate warm-up and decay strategy used in our experiments

1×10^{-7} for ACDC, Hippocampus, and Spleen datasets, 5×10^{-7} for PROMISE12, and 2×10^{-6} for MMWHS, with N being set to 300 for all datasets. We fixed `max_epoch` to 75 for ACDC dataset and to 80 for the remaining datasets.

The generation of mini-batches is crucial for the contrastive learning. Following Chaitanya *et al.* (2020), we randomly sampled 10 scans from ACDC and drew one image per partition for each scan, resulting in 30 images per iteration. However, for the two other datasets, we randomly sampled 30 scans without considering meta-labels. We used a supervised loss $\mathcal{L}$sup to guide the network optimization during both the fine-tune procedure (for evaluating the quality of the pre-trained weights) and the semi-supervised training. Although different loss functions can be considered, such as the Dice loss (Zhao *et al.*, 2020b) and Tversky loss (Salehi, Erdogmus & Gholipour, 2017), we adopted the well-known cross entropy loss in all experiments. This loss is defined as

$$\mathcal{L}_{\text{sup}} = -\frac{1}{|\mathcal{D}_l|\,|\Omega|}\sum_{(\mathbf{x},\mathbf{y})\in\mathcal{D}_l}\sum_{i\in\Omega}\sum_{c=1}^{C} y_{ic}\log p_{ic}(\mathbf{x}) \qquad \text{(A III-2)}$$

where $\mathcal{D}_l$ is labeled dataset, Ω is the 2-D pixel space and $p_{ic}(\mathbf{x})$ is the probability of class $c \in \{1,\ldots,C\}$ predicted by the network at pixel i .

For our contrastive pre-training experiments, we take the feature maps $E(\mathbf{x})$ at the end of the encoder and, following Chen *et al.* (2020a); Chaitanya *et al.* (2020), project them to low-dimensional vectors via a projector g which consists of an average pooling to flatten the output of the encoder followed by a two-layer MLP to have a final dimensionality of 256.

Our proposed approach and compared methods have several tunable hyper-parameters, such as the weighting coefficient for each loss, the evolution strategy for γ from Equ. (16) of the main paper, etc. Those hyper-parameters were chosen based on the performance of the method on the validation set. We ran our experiments on a computing cluster with Nvida P-100 GPUs. Running the pre-training stage usually takes less than 4 hours for all datasets, while the semi-supervised learning can take $4-6$ hours depending on the approach.

For our experiments shown in Tables 1, 2 and 3 of the main paper, we fixed the hyper-parameters as follows. The temperature τ was set to 0.07, close to related works (Chen *et al.*, 2020a; Wang & Liu, 2021; Chaitanya *et al.*, 2020) and working fine across different datasets. For λ in Equ. (14), we used $\lambda_1 = 1.0$, $\lambda_2 = 1{\times}10^{-3}$ and $\lambda_3 = 1.0 \times 10^{-2}$ for the ACDC dataset, while values of $\lambda_1 = 1.0$, $\lambda_2 = 1 \times 10^{-1}$ were selected for the PROMISE12 and MMWHS dataset. For our combined model of Equ. (15), which involves a supervised loss, a regularization loss (Mean Teacher) and our SP-based contrastive loss, we fixed λ_{reg} to 0.1 and λ_{sp} to $5{\times}10^{-3}$ for the ACDC dataset, whereas λ_{reg} was set to 0.2 and λ_{sp} to $2{\times}10^{-2}$ for both the Prostate and MMWHS datasets. We also include a hyper-parameter sensitivity analysis of λ in Sec. 5.6 to show the contribution of each loss component.

4.3 Details of compared methods

In Table 2 and 3, we compare our proposed self-paced contrastive method against several baselines, ablation variants of our method, and recently-proposed approaches for semi-supervised medical segmentation. We give a description of tested approaches below:

- **Contrastive loss (Chaitanya *et al.*, 2020):** This setup evaluates the performance of a normal contrastive learning loss $\mathcal{L}\text{con}$ as in Equ. (2) of the main paper. In this case, we only consider the encoder features without our self-paced learning strategy to adapt the importance of positive pairs in the contrastive loss.
- **Self-Paced Contrastive Loss (SP-Con):** This is our full formulation of Equ. (3) in the main paper which exploits meta-labels for computing $\mathcal{L}_{\text{SP}}{}^{k}$, where $k \in \{1, \ldots, K\}$. As explained in Section 5.2, this loss can be used for pre-training as well as for semi-supervised learning. In our experiments, we test different combinations of this loss and other semi-supervised approaches.
- **Mean Teacher (Tarvainen & Valpola, 2017):** This powerful method for semi-supervised learning adopts a teacher-student framework, where two networks sharing the same architecture learn from each other. Given an unlabeled image $\mathbf{x}$, the student model $p^s(\cdot)$ seeks to minimize the prediction difference with the teacher network $p^t(\cdot)$ whose weights are a temporal exponential moving average (EMA) of the student's:

  $$\mathcal{L}_{\text{MT}} = \frac{1}{|\mathcal{D}_u|\,|\Omega|} \sum_{\mathbf{x}\in\mathcal{D}_u} \sum_{i\in\Omega} \sum_{c=1}^{C} (p^t_{ic}(\mathbf{x}) - p^s_{ic}(\mathbf{x}))^2 \qquad \text{(A III-3)}$$

  Following the standard practice, we fix the decay coefficient to 0.999. The coefficient balancing the supervised and regularization losses is selected by grid search, from 1×10^{-4} to 10.0.
- **Entropy Minimization (Entropy Min.) (Vu *et al.*, 2019):** This method, which has been successfully applied in semi-supervised classification (Grandvalet & Bengio, 2005) and segmentation (Vu *et al.*, 2019) with domain gap, imposes a low conditional entropy on unlabeled images:

  $$\mathcal{L}_{\text{ent}} = -\frac{1}{|\mathcal{D}_u|\,|\Omega|} \sum_{\mathbf{x}\in\mathcal{D}_u} \sum_{i\in\Omega} \sum_{c=1}^{C} p_{ic}(\mathbf{x}) \log p_{ic}(\mathbf{x}). \qquad \text{(A III-4)}$$

  By increasing its confidence for unlabeled images, the network pushes the decision boundary away from dense regions of the input space, thereby improving generalization. For this method, we performed a hyper-parameter search on the coefficient balancing the two losses, $\mathcal{L}_{\text{sup}}$ and $\mathcal{L}_{\text{ent}}$, from 1×10^{-5} to 1×10^{-1}.

- **MixUp (Zhang *et al.*, 2017b):** We also evaluated the effectiveness of mixup, an effective data argumentation strategy on medical image segmentation, following Chaitanya *et al.* (2020).
- **Adversarial Training (Zhang *et al.*, 2017c):** This semi-supervised segmentation method trains a segmentation network and a classifier-based discriminator jointly in a min-max game. Its core idea is to enforce the segmentation predictions on unlabeled images being indistinguishable from those of labeled images, thus aligning the output distributions between labeled and unseen images. This method works particularly well in a scenario where scans present large variability causing a domain gap.
- **Discrete Mutual Information maximization (Peng *et al.*, 2021a):** This semi-supervised segmentation technique maximizes the mutual information between two sets of feature maps undergoing different geometric transformations. These feature maps are taken from different hierarchical levels of the segmentation network, thus regularizing the network at multiple scales. It was shown effective for segmenting medical images with limited supervision. In this experiment, we optimize the mutual information between features taken from both the encoder (as our proposed method), and from the decoder.
- **Global and Local Contrastive (Chaitanya *et al.*, 2020):** This last approach is our full implementation of (Chaitanya *et al.*, 2020), which takes into account not only the encoder's features as a global descriptor, using Equ. (2) of the main paper, but also dense features from decoder blocks that allow an effective contrastive learning at the pixel level. For the decoder, we chose the output of the third decoder block and resized the dense features to a fixed resolution of 10×10 pixels with adaptive average pooling. For each feature map, we then randomly sampled 5 different spatial locations and performed contrastive learning using Equ. (1) of the main paper on these $5 \times 2N_{\text{batch}}$ vectors, where N_{batch} is the number of images in the current

batch. Following Chaitanya *et al.* (2020), we adopted a two-step pre-training strategy, and pre-trained the decoder while freezing the encoder. Using the well pre-trained weights, we report the test score when fine-tuning models on a few labeled scans.

5. Additional experimental results

5.1 Supplementary experiments on two extra datasets

The proposed method could work with any segmentation task where meta-labels are available. This includes segmenting volumetric data of any organ for which a rough correspondence can be obtained between 2D slices in the volume. In order to further highlight the robustness of our proposed method, we carried out additional experiments on two extra datasets segmenting the hippocampus and spleen from MRI and CT images. For these two dataset, we only tested our proposed variant SP-Con (*pre-train*) on the slice position meta-label ($\mathcal{L}^1_{\text{sp}}$) and compared it against strong concurrent approaches such as Contrastive ($\mathcal{L}^1_{\text{con}}$) (Chaitanya *et al.*, 2020) and Mean Teacher (Tarvainen & Valpola, 2017).

As one can see from Table III-1, our proposed method outperforms Contrastive in most cases and reaches a performance comparable with Mean Teacher. This confirms the general usefulness of our self-paced learning strategy for the semi-supervised segmentation of different organs.

Table-A III-1 3D DSC performance of the proposed SP-Con (*pre-train*) variant using slice position as meta-label and other approaches on hippocampus and spleen datasets with few labeled scans.

Method	Pretrain	Train	Hippocampus			Spleen	
			1 scan	**2 scans**	**4 scans**	**2 scan**	**4 scans**
Baseline	-	$\mathcal{L}_{\text{sup}}$	60.87	73.33	78.82	65.51	68.59
Mean Teacher (Perone & Cohen-Adad, 2018)	-	$\mathcal{L}_{\text{sup}} + \mathcal{L}_{\text{mt}}$	**70.06**	75.65	80.60	55.02	68.10
Contrastive (Chaitanya *et al.*, 2020)	$\mathcal{L}^1_{\text{con}}$	$\mathcal{L}_{\text{sup}}$	64.40	75.00	**81.45**	65.14	67.21
SP-Con (*Pre-train*)	$\mathcal{L}^1_{\text{sp}}$	$\mathcal{L}_{\text{sup}}$	66.70	**76.89**	81.25	**69.05**	**69.64**

5.2 Fine-tuning the model on the whole dataset with annotations

We showed previously that pre-training a network with images and meta-labels from the dataset helps to improve performance when fine-tuning it using a few labeled examples. In this section, we consider the following question: "is our pre-training loss helpful when fine-tuning the model on the entire dataset with ground-truth annotations, instead of just a few of them?" To answer this question, we first pre-trained a network with our SP-Con (*pre-train*) loss and fine-tuned it with different numbers of labeled images on the ACDC dataset. Table III-2 summarizes the segmentation performance of our method after fine-tuning, and compares it to Mean Teacher.

Table-A III-2 3D DSC improvements brought by SP-Con (Pre-train) with slide position as meta-label with various numbers of labeled scans.

Methods	**1 scan**	**2 scans**	**4 scans**	**8 scans**	**175 scans**
Baseline	57.53	67.06	75.64	82.64	88.06
Mean Teacher	62.85	72.84	79.12	84.35	N/A[2]
SP-Con (*pre-train*)	73.99	81.01	82.83	84.29	88.35
Our Improvement	16.46	13.95	7.18	1.65	0.29

We observe that our method's relative improvement with respect to the baseline reduces as the number of labeled samples increases. This is expected since there is no additional unlabeled data to exploit when using the entire set of images as labeled data (175 scans). However, SP-Con still yields a small improvement (0.29%) compared to the full supervision (Baseline with 175 scans as labeled images).

5.3 Segmentation performance when randomly selecting labeled volumes per experiment

For the previous experiments, when fine-tuning the model or using a semi-supervised training with a few labeled data, we kept a fixed split on the labeled/unlabeled images for each number

[2] Mean Teacher requires unlabeled examples for its consistency loss, thus was not considered for the full supervision setting.

of labeled scan, and reported results were obtained by averaging performance from three independent runs with different random seeds controlling model initialization, data augmentation randomness, and data fetch ordering. In order to remove the possible bias from the choice of these fixed splits, we ran three additional experiments using different labeled/unlabeled splits per experiment, the results of which are reported in Table III-3.

Table-A III-3 3D DSC performances for different splits on the ACDC dataset. Best cases are highlighted in **bold**.

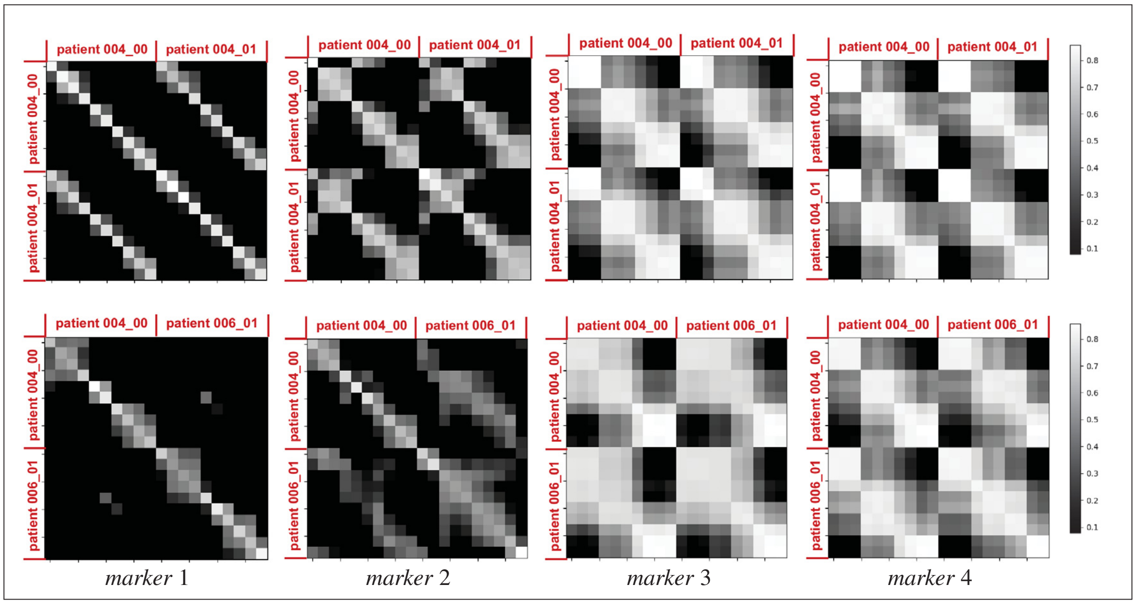
Method	Pretrain	Train	Split 1		Split 2		Split 3		Mean (std) over splits	
			1 scan	2 scans	1 scan	2 scans	1 scan	2 scans	1 scan	2 scans
Baseline	-	$\mathcal{L}_{\text{sup}}$	61.36	71.11	35.76	71.48	39.74	76.44	45.62 (11.25)	73.01 (2.43)
Unsup. Con.	$\mathcal{L}_{\text{con}}^{\text{unsup.}}$	$\mathcal{L}_{\text{sup}}$	65.70	77.84	50.69	72.87	59.65	80.22	58.68 (6.17)	76.98 (3.06)
Unsup. Con. + SP	$\mathcal{L}_{\text{sp}}^{\text{unsup.}}$	$\mathcal{L}_{\text{sup}}$	69.34	77.49	61.96	73.11	66.55	80.82	65.95 (3.04)	77.14 (3.16)
Contrastive	$\mathcal{L}_{\text{con}}^{1}$	$\mathcal{L}_{\text{sup}}$	72.56	**80.58**	67.68	73.97	68.64	82.35	69.63 (2.11)	78.97 (3.61)
	$\mathcal{L}_{\text{con}}^{2}$		68.19	76.69	43.69	71.57	57.34	80.31	56.41 (10.02)	76.19 (3.59)
	$\mathcal{L}_{\text{con}}^{3}$		65.40	76.33	64.93	69.10	58.82	80.06	63.05 (3.00)	75.16 (4.55)
Mean Teacher	-	$\mathcal{L}_{\text{mt}}$	73.04	78.97	60.91	72.82	55.07	80.26	63.01 (7.48)	77.35 (3.25)
SP-Con (*pre-train*)	$\mathcal{L}_{\text{sp}}^{1}$	$\mathcal{L}_{\text{sup}}$	**76.24**	79.96	**68.18**	**76.46**	**74.18**	**82.46**	**72.87 (3.42)**	**79.63 (2.46)**
	$\mathcal{L}_{\text{sp}}^{2}$		71.77	80.08	58.07	71.95	58.02	81.07	62.62 (6.47)	77.70 (4.09)
	$\mathcal{L}_{\text{sp}}^{3}$		66.77	77.10	63.65	73.41	62.03	82.11	64.15 (1.97)	77.54 (3.57)

In this new set of experiments, we followed the same experimental protocols as in previous one, where we first pre-train the network with different contrastive losses (unsupervised, contrastive with three meta-labels, and those with self-paced learning) and then fine-tune them using a few labeled scans. However, this time vary the scans used as labeled data for each run and compute the final performance by averaging results from the three experiments.

As one can see from Table III-3, using different splits for the labeled data leads to a large variance in performance. When given a single labeled scan, the baseline DSC for different splits varies from 35.75% to 61.36%, with a standard derivation up to 11.25%. On the other hand, our model using meta-labels obtains more robust results (e.g., standard deviation of 3.42% for $\mathcal{L}_{\text{sup}}^{1}$ using one labeled scan) and outperforms the contrastive counterparts in all but one cases.

5.4 Self-paced learning analysis

As discussed in Section 5.1 of the main paper, we include the self-paced weights w_{ij} taken from different epochs, corresponding to four different γ (marked as orange star) in Fig. 1 of the main paper. Notice that w_{ij} is only defined on positive pairs, however we visualize w_{ij} for all image pairs to verify whether our proposed self-paced strategy can successfully learn weights that capture the corresponding meta-label's quality. We train our proposed loss using slice position as the only meta-label (i.e., $\mathcal{L}_{\text{sp}}^{1}$).



marker 1 *marker* 2 *marker* 3 *marker* 4

Figure-A III-6 Self-paced importance weight w_{ij} for two scans during the optimization. We plot not only the w_{ij} for positive pairs, but also for negative pairs.

In Fig. III-6, we first compare two scans from the same patient with different cardiac phases (first row of the figure). As can be seen, in the beginning of the training (*marker* 1), self-paced weights have a high value for pairs of slices from the same position, even when corresponding to different phases of the cardiac cycle. As the training progresses, our proposed strategy learns to assign relatively high values on positive pairs, and keep low values for negative pairs (*marker* 2). At the end of the training, sharper blocks form in weight matrix based on meta-label classes, and the network learns to distinguish different slice partitions (*marker* 3-4). In contrast, conventional contrastive loss lacks the ability to learn gradually and does not have an effective selection mechanism in the beginning of training. We also conducted a similar analysis with two scans taken from different patients (second row of the figure). It can be seen that the proposed loss learns to assign weights based on the network's uncertainty for each pair. We can observe that the network is slower to assign high values to positive pairs across patients in the first few epochs, since they differ in terms of appearance.

5.5 Sensitivity to γ parameters

The next experiment investigates the impact of γ_{start} and γ_{end} in Equ. (16) of the main paper, where we fixed $p = 1/2$ and adopt the linear strategy to ramp-up γ. Theorem 2 of the main paper states that ℓ_{ij} is bounded by $\log 2N + 2/\tau$. However, for the linear strategy, where $w_{ij} = 1 - \frac{1}{\gamma}\ell_{ij}$, we found that further increasing γ boosts the segmentation performance. Table III-4 shows the results of a grid search varying these two hyper-parameters (γ_{start} in rows and γ_{end} in columns) and measuring the performance on the validation set. As can be seen, the segmentation DSC reaches its maximum value when γ_{start} is 2.0. Given a fixed ramp-up strategy, choosing a small γ_{start} causes the learning to ignore most of the examples in the beginning of the training and, thus, results in under-fitting. In contrast, a large value of γ_{start} pushes the network to treat all examples equally. Since γ_{end} controls the level of weights in the end of the training, it has a similar behavior as γ_{start}.

5.6 Sensitivity to λ_k parameters

We then present a sensitivity analysis for λ_k in Equ. (14) on the ACDC dataset. As $\mathcal{L}\text{sp}^1$ (the one using slice partition as the supervision signal) resulted in the best performance across all dataset, we fixed λ_1 as 1.0 and performed a grid search on λ_2 and λ_3 with logarithmic scales ranging from 0.1 to 0.001. We report the 3D DSC performance on the validation set using just

Table-A III-4 Sensitivity analysis of hyper-parameter γ.

γ_{start}	γ_{end}				
	40.0	50.0	60.0	70.0	80.0
1.0	72.65	73.83	72.98	72.60	72.99
1.5	72.56	73.95	74.97	72.56	73.79
2.0	73.20	**75.08**	73.32	73.39	72.70
2.5	72.61	73.75	72.26	72.49	73.20
3.0	72.95	73.50	72.89	73.96	73.33
3.5	73.62	73.08	74.28	74.38	73.06

one scan as labeled data, as shown in Table III-5. We see that varying both λ_2 and λ_3 leads to relatively stable performance for the downstream segmentation tasks, while the best performance is achieved by setting $\lambda_2 = 0.01$ and $\lambda_3 = 0.001$. This suggests that, for the ACDC dataset, using Patient Identity as the meta-label brings a more useful information compared to Cardiac cycle phase.

Table-A III-5 Sensitivity analysis of hyper-parameters λ_2 and λ_3 for the ACDC dataset.

λ_2	λ_3		
	0.1	0.01	0.001
0.1	73.09	74.12	74.61
0.01	73.99	74.56	**74.89**
0.001	73.52	74.38	74.52

5.7 Visual result inspection

Last, we provide a visual inspection on segmentation predictions for different approaches on the three first datasets. As can be seen, the proposed method effectively enhances the segmentation quality on the different datasets, and reaches a segmentation prediction closer to the ground truth than other tested approaches.

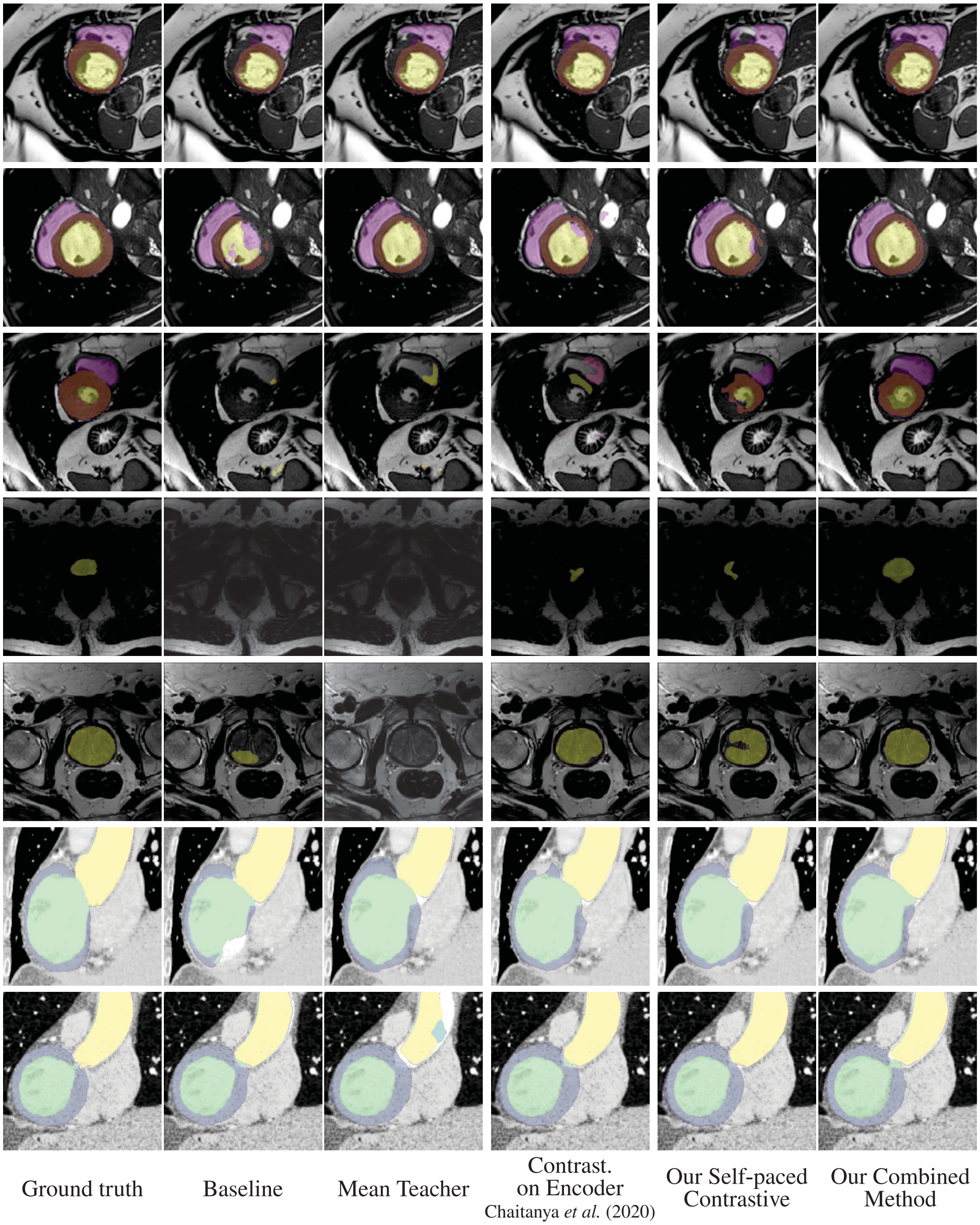

Ground truth
Baseline
Mean Teacher
Contrast. on Encoder Chaitanya *et al.* (2020)
Our Self-paced Contrastive
Our Combined Method

Figure-A III-7 Visual comparison of tested methods on test images. **Rows 1–3**: the ACDC; **Rows 4–5**: the PROMISE12; **Rows 6–7**: MMWHS.

APPENDIX IV

APPENDIX FOR PAPER «BOUNDARY-AWARE INFORMATION MAXIMIZATION FOR SELF-SUPERVISED MEDICAL IMAGE SEGMENTATION»

1. Presentation

This appendix provides conceptual diagram of our proposed method, detailed experimental setup, dataset description, data pre-processing steps, and complementary experimental results for Chapter 6 – "Boundary-aware Information Maximization for Self-supervised Medical Image Segmentation", which was submitted to ICML 2022 conference.

2. Diagram

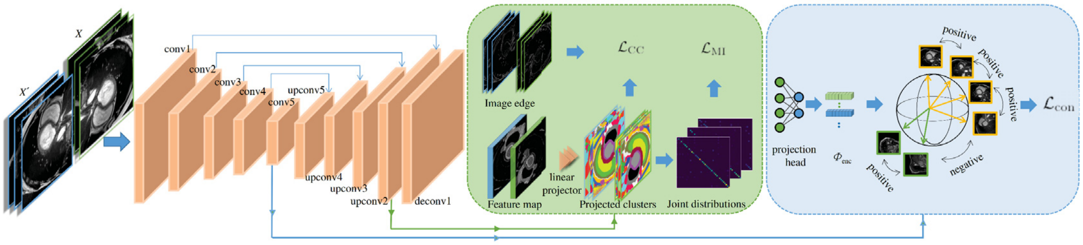


Figure-A IV-1 Schematic Diagram of our proposed method. Our method consists of three individual objectives. *Global* contrastive loss $\mathcal{L}_{\text{con}}$ enforces images with similar anatomical structures to be pull close, providing global context information for the encoder. Our MI-based loss $\mathcal{L}_{\text{MI}}$ aims to cluster the dense local embeddings into K balanced and confident classes, while our boundary-aware loss $\mathcal{L}_{\text{CC}}$ aligns the boundaries of these clusters to image edges.

3. Related work

Inspired by the recent success of representation learning, various approaches based on pre-training have been investigated for segmentation. These approaches seek to acquire a discriminative image representation from unlabeled data, in an independent pre-train stage. In medical image segmentation, Chen *et al.* (2019a); Bai *et al.* (2019) proposed to predict the relative position of patches in MR images. Taleb *et al.* (2021) extended the jigsaw puzzle solving pretext task to images from multiple MRI modalities. Zhou *et al.* (2021) proposed Models Genesis, a denoising auto-encoder that reconstructs an MR image given its degraded version as input. Contrastive learning has also shown promising results to boost the performance of downstream tasks using unlabeled images. In this approach, a network is pre-trained to bring closer the feature embeddings of an image under different transformations (positive pairs), while pushing away those from different images (negative pairs). This idea was used to learn a global representation at the end of the network's encoder, using meta-labels on anatomical similarity or subject ID to define the positive pairs (Chaitanya *et al.*, 2020; Peng *et al.*, 2021b; Zeng *et al.*, 2021). To also pre-train the decoder, the method in Chaitanya *et al.* (2020) defined positive or negative embedding pairs based on their spatial distance in a feature map, those with a large distance considered as negative while those at the same spatial position but coming from different transformations as positives. Hu *et al.* (2021) proposed using small set of pixel-wise annotations to guide the learning of dense features in pre-training. The feature embeddings of pixels with the same label are considered as positive pairs and are then clustered together by the contrastive loss. While this guided approach helps learn a better local representation, it requires manual annotations and therefore it is not unsupervised. Ouyang *et al.* (2020) instead employed superpixels for the contrastive objective, however their approach is defined in the context of few-shot segmentation.

Clustering has also been used to pre-train a network with unlabeled images (Caron *et al.*, 2020; Ji *et al.*, 2019; Cho, Mall, Bala & Hariharan, 2021; Fang, Liang, Shao, Dong & Li, 2021). Surprisingly, only a few papers have explored this self-supervised learning approach for medical image segmentation (Peng *et al.*, 2021a; Ahn, Feng & Kim, 2021). Our method extends the IIC deep clustering approach (Ji *et al.*, 2019) with an improved loss that encourages clusters to be consistent across different transformations and follow the region boundaries in the image.

Accurately predicting the boundaries of anatomical structures, tissues or lesions is essential for medical image segmentation, and boundary-aware training methods have been widely explored. To achieve this goal, most methods use a multi-task learning strategy with a secondary loss function focusing on boundary information (Li *et al.*, 2020b; Xue *et al.*, 2020; Shen, Wang, Zhang & McKenna, 2017). Another approach adopts a discriminator to embed the ground-truth boundary (Wei, Shi, Song, Ji & Han, 2020). Unlike our boundary-aware information maximization method for unsupervised representation learning, these approaches require annotated images and thus are limited to supervised or semi-supervised settings.

4. Proof of Proposition 6.3.1

Proof. For the lower bound, we use the inequality $\mathcal{H}(P, Q) = (P \,\|\, Q) + \mathcal{H}(P)$ to obtain

$$\begin{aligned}
\mathcal{H}(\frac{1}{K}\mathrm{I}_K, \mathcal{P}_{\text{joint}}) &= \underbrace{(\frac{1}{K}\mathrm{I}_K \,\|\, \mathcal{P}_{\text{joint}})}_{\geq 0} + \mathcal{H}(\frac{1}{K}\mathrm{I}_K) && \text{(A IV-1)}\\
&\geq -\sum_{j=1}^{K}\sum_{k=1}^{K}\frac{1}{K}1[j=k]\log(\frac{1}{K}1[j=k]) && \text{(A IV-2)}\\
&= -\sum_{k=1}^{K}\frac{1}{K}\log\frac{1}{K} = \log K. && \text{(A IV-3)}
\end{aligned}$$

For the upper-bound, we use the Jensen inequality to get

$$\begin{aligned}
\mathcal{H}(\frac{1}{K}\mathrm{I}_K, \mathcal{P}_{\text{joint}}) &= \sum_{j=1}^{K}\sum_{k=1}^{K}\frac{1}{K}1[j=k]\log\mathcal{P}_{\text{joint}}^{(j,k)} && \text{(A IV-4)}\\
&= -\sum_{k=1}^{K}\frac{1}{K}\log\left(\frac{1}{N}\sum_{i=1}^{N}\widehat{p}_{ik}\,\widetilde{p}_{ik}\right) && \text{(A IV-5)}\\
&\leq -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\frac{1}{K}(\log\widehat{p}_{ik} + \log\widetilde{p}_{ik}) && \text{(A IV-6)}\\
&= \frac{1}{N}\sum_{i=1}^{N}\mathcal{H}(\mathbf{u}, \widehat{\mathbf{p}}_i) + \mathcal{H}(\mathbf{u}, \widetilde{\mathbf{p}}_i) \quad \square && \text{(A IV-7)}
\end{aligned}$$

□

5. Datasets

We assess the performance of our proposed method and compare it with other SOTA approaches on two clinically-relevant datasets: the automatic cardiac diagnosis challenge (ACDC) and the Prostate MR image segmentation 2012 challenge (Promise12). These two datasets cover different anatomical structures, present different acquisition resolutions, and are widely used to verify the effectiveness of semi-supervised segmentation algorithms.

ACDC dataset: The ACDC dataset[1] consists of 200 short-axis cine-MRI scans from 100 patients, evenly distributed in 5 subgroups: normal, myocardial infarction, dilated cardiomyopathy, hypertrophic cardiomyopathy, and abnormal right ventricles. For each patient, two annotated scans correspond to end-diastolic (ED) and end-systolic (ES) phases are provided, which 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 regions of interest: left ventricle endocardium (LV), left ventricle myocardium (Myo), right ventricle endocardium (RV) are labeled from background and delineated pixel-wisely by human experts. We consider the 3D-MRI scans as 2D images through-plane due to the high anisotropic acquisition resolution, and re-sample them to a fix space ranging of 1.0 × 1.0 mm. Following Peng *et al.* (2021b), we normalize the pixel intensities based on the 1% and 99% percentile of the intensity histogram for each scan. Normalized slices are then cropped to 384 × 384 pixels, coarsely centered based on the foreground delineation of the ground truth. We select slices from 175 random scans as our training set, from which we again randomly select 1, 2 or 4 scans[2] as our labeled data, representing representing 0.5% to 2% of all available data, and consider others as unlabeled data.

[1] Publicly-available by https://www.creatis.insa-lyon.fr/Challenge/acdc/index.html

[2] These labeled splits were also kept untouched for experiments employing *pre-train* and *fine-tune* strategies, as well as those relying on semi-supervised losses.

We then randomly divide the remaining 25 scans into a validation set and a test set, comprised of 8 scans and 17 scans respectively. Both the validation and test sets were set aside during model optimization. For experiments with *pre-train* and *fine-tune* strategies, we use all *training* data without any pixel-wise annotation for *pre-train* and evaluate the representation ability by fine-tuning the obtained network on a few labeled scans via a cross-entropy loss. We also used various data augmentation transformations $\mathcal{T}(\cdot)$ for both labeled and unlabeled images, including random crops of 224 × 224 pixels, random flip, random rotation, and color jitter from the `Pillow` library. It is also worthy to notice that the main experimental results from this dataset were obtained by considering three-class segmentation as three binary segmentation tasks.

Promise12 dataset: Our second dataset[3] focuses on prostate segmentation and is composed of multi-centric transversal T2-weighted MR images from 50 subjects. These images were acquired from different vendors and with various acquisition protocols, and are thus representative of typical MR images acquired in a clinical setting. Image resolution ranges from 15 × 256 × 256 voxels to $54 \times 512 \times 512$ voxels with a spacing ranging from $2 \times 0.27 \times 0.27$ mm to $4 \times 0.75 \times 0.75$ mm. Also in this case, we slice these volumetric images into 2D images along the short-axis and resized them to a resolution of 256×256 pixels. We equally performed a normalization on pixel intensities based on 1% and 99% percentile for each scan. We randomly selected 40 scans as training data, 3 scans for validation, and 7 scans for testing. To test methods in annotation-scarcity regime, we chose 4, 6, and 8 scans from these training examples as the labeled images, while keeping others as unlabeled. We also employed rich data transformation prior to $\mathcal{T}(\cdot)$. These `Pillow`-based transformations include random crop of 224×224 pixels, random flip, random rotation within a range of $[-10°, 10°]$, and color jitter.

[3] Publicly-available by https://promise12.grand-challenge.org/

6. Detailed comparison methods

We implement various state-of-the-art methods for comparison, including:

Contrastive-based method (Chaitanya *et al.*, 2020): As our closest work, this approach acquires discriminative representation from unlabeled images via two *global* and *local* contrastive learning. *Global* contrastive learning pre-trains the encoder of the segmentation network to distinguish the global context information such as the anatomical similarities between two slices, while *local* contrastive learning focuses on dense embeddings and enforces pixels undergoing different transformations to be close and pixels at different spatial locations to be pushed away. We refer *Contrast (Enc)* as our `PyTorch` re-implementation of the variant using only *global* contrastive objective, *Contrast (Dec)* as the variant employing only *local* contrastive learning, and *Contrast (Enc+Dec)* as the full implementation using *both* contrastive objectives. For *Contrast (Enc)* variants, we pre-train the encoder of the segmentation network to distinguish whether two slices comes from similar slice position by assuming the volumetric scans are coarsely aligned. Towards this goal, we manually split an ACDC scans into three partitions, while we fixed the partition numbers for PROMISE12 as 5, similar to Peng *et al.* (2021b). A nonlinear projector is used to convert the global representation to representation vector, comprised of an average pooling layer, 2-layer MLP with LeakyReLU as the activation, and a normalization layer. For the variants using *local* contrastive learning, we take the dense embeddings from the layer before last 1×1 convolutions. These dense embeddings are then projected to pxiel-wise vectors by a dense projector, consisting of an adaptive average pooling of size 20×20 to reduce the spatial size, 2-layer MLP with 1×1 convolutions with LeakyReLU as the activation, and a normalization layer. Positive and negative pairs are then defined on these 20×20 grid, similar to Chaitanya *et al.* (2020). It is worthy to point out that we only optimize the parameters for the encoder in the *pre-train* stage only when *global* contrastive loss is used, while the whole network except last 1×1 convolution is optimized when *local* contrastive loss is employed. We employ two stage strategy: *pre-train* and *fine-tune* to evaluate the quality of the learned representation.

IIC (Dec) (Ji *et al.*, 2019): This is the original method proposed in (Ji *et al.*, 2019) for unsupervised image clustering, which corresponds to our optimization objective $\mathcal{L}_{\text{MI}}$ with $\alpha = 0.0$. This method has being successfully used in image clustering, as well as for unsupervised natural image segmentation, but it can only work well with coarse classes. We follow the exact protocol and hyper-parameters as our proposed method and report the 3D DSC score on the test set.

IMSAT (Dec) (Hu *et al.*, 2017a): This method seeks to maximize MI over categorical distributions from dense embeddings in a similar but different formulation: $I = I(X, p(X))$, where X is the image set while $p(X)$ is the cluster assignment distribution given X. This objective has been successfully applied in image clustering (Hu *et al.*, 2017a). We adapt this loss for dense embedding clustering and this method shares the same experimental protocols as our proposed method. We equally evaluate its performance using *pre-train* and *fine-tune* strategy.

Entropy Minimization (EM) (Vu *et al.*, 2019): This method has been successfully applied in semi-supervised classification and segmentation with domain gap, and imposes a low conditional entropy on unlabeled images: $\mathcal{L}_{\text{ent}} = -\frac{1}{|\mathcal{D}_u||\Omega|} \sum_{x \in \mathcal{D}_u} \sum_{i \in \Omega} p_i(x) \log(p_i(x))$. By increasing its confidence for unlabeled images, the network pushes the decision boundary away from dense regions of the input space, therefore improving generalization. For this method, we performed a hyper-parameter search on the coefficient balancing the cross-entropy and $\mathcal{L}_{\text{ent}}$, from 1×10^{-4} to 1.0. We evaluate this method in a standard semi-supervised setting with randomly initialized network parameters.

MixUp (Zhang *et al.*, 2017b): We also evaluated the effectiveness of mixup, an effective data argumentation strategy on medical image segmentation, following Chaitanya *et al.* (2020). In this method, we interpolate two labeled images with an index sampled from $Beta(\alpha, \alpha)$ distribution and enforce the network to output the prediction as the interpolation of the two annotations. We fix $\alpha = 1$ and the coefficient weighting the mixup loss is selected by grid search from 1×10^{-5} to 0.1.

Mean Teacher (MT) (Perone & Cohen-Adad, 2018): This semi-supervised segmentation method adopts a teacher-student framework, in which two networks sharing the same architecture learn from each other. Given an unlabeled image x, the student model $p^s(\cdot)$ seeks to minimize the prediction difference with the teacher network $p^t(\cdot)$, whose weights are a temporal exponential moving average (EMA) of the student's: $\mathcal{L}_{\mathrm{MT}} = -\frac{1}{|\mathcal{D}_u||\Omega|} \sum_{x \in \mathcal{D}_u} \sum_{i \in \Omega} |p_i^s(x) - p_i^t(x)|^2$. We fix the decay coefficient to 0.99. The coefficient balancing the supervised and regularization losses is selected by grid search, from 1×10^{-4} to 10.

Uncertainty-aware Mean Teacher (UA-MT) (Yu *et al.*, 2019): This semi-supervised approach introduces uncertainty for teacher network, which is achieved by Monte-Carlo dropout through multiple inferences. In our implementation, we forward through the teacher network unlabeled images four times and the uncertainty is obtained by computing the pixel-wise entropy of these predictions. We then use a linearly increased threshold T, ranging from $\frac{3}{4} \times \log(K)$ to $\log(K)$ to exclude from $\mathcal{L}_{\mathrm{MT}}$ pixels having high uncertainty. We keep other settings the same as our Mean Teacher method and evaluate method's performance in a standard semi-supervised setting.

Interpolation Consistency Training (ICT) (Verma *et al.*, 2019): The next method we tested applies mixup method with teacher-student framework. In this approach, interpolated images are obtained by mixing up two unlabeled images. The student network is encouraged to output the prediction as the interpolation of their predictions given by the teacher network. We follow (Verma *et al.*, 2019) to set α as 0.1 and again grid search the weighting coefficient for the regularization objective, from 1×10^{-5} to 0.1.

Adversarial training(AT) (Zhang *et al.*, 2017c): Our last method trains a segmentation network and a classifier-based discriminator jointly in a min-max game. The core idea is to enforce the segmentation predictions on unlabeled images being indistinguishable from those of labeled images, thus aligning the output distributions between labeled and unseen images. This method works particularly well in a scenario where the image scans present large variability causing a domain gap. We evaluate this method in a standard semi-supervised setting and grid-search the regularization coefficient, from 1×10^{-6} to 0.1.

7. Implementation details

Network Architecture: We used U-Net (Ronneberger *et al.*, 2015) as our main network architecture, which consists of five symmetric blocks of encoder and decoder. As shown in Fig. IV-1, we assign different names to these blocks and our global embeddings and dense embeddings are taken from `conv5` and `upconv2`. A first nonlinear projector is used to convert the global representation to representation vector, comprised of an average pooling layer, 2 MLP layers with LeakyReLU as the activation, followed by a normalization layer. In contrast, we simply employ a linear projector, including 1×1 convolution followed by a K-way softmax for the dense embeddings. Learnable parameters are optimized using stochastic gradient descent (SGD) with a RAdam Optimizer (Liu *et al.*, 2019).

Training hyper-parameters: Our main experiments adopt the two-stage training strategy: pre-training the whole network on all training data without labels and fine-tune it with a few labeled scan. For both stages, we employed a learning rate decay strategy, where the initial learning rate lr is increased N times in the first 10 epochs, followed by a cosine decay strategy for the rest N_{epoch} training epochs. We set $lr = 5 \times 10^{-7}$, $N = 400$, and $N_{\text{epoch}} = 50$ for the ACDC in pre-train stage, $lr = 1 \times 10^{-7}$, $N = 200$, and $N_{\text{epoch}} = 50$ for ACDC in fine-tune stage. As for the Promise12 dataset, we simply modify lr to 1×10^{-6} for the fine-tune stage. We define an epoch in our experiments as the N update iterations, within which images are randomly selected from their respective dataset with replacement. For ACDC, we fixed N as 200 iterations while for

PROMISE12, we increase N to 400. For concurrent methods employing semi-supervised setting, we follow exactly the same configuration as adopted in fine-tune stage. As shown in Equ. (A II-4), our method requires only one weighting coefficient which balances the importance of our $\mathcal{L}_{\text{CC}}$ and we simply set it to 1.0 for both datasets.

Details on the transformation $\mathcal{T}(\cdot)$: Our proposed method heavily relies on $\mathcal{T}(\cdot)$ to create transformation equivalent pairs of cluster distribution: $\widehat{\mathbf{p}} = g(s(\mathcal{T}(x)))$ and $\widetilde{\mathbf{p}} = g(\mathcal{T}(s(x)))$. We set $\mathcal{T}(\cdot)$ as the cascade of intensity transformations and geometric transformations. When $\mathcal{T}(\cdot)$ takes an input x as the raw image, we apply gamma correction within a range of [0.5, 2.0], as well as a set of random affine transformation, consisting of random scale within a range of [0.8, 1.3], random rotation within a range of [−45°, 45°], and random translation within a range of [−10%, 10%]. Whereas when $\mathcal{T}(\cdot)$ takes the input as the embedding s of the image x, we ignore the intensity transformation and apply only the random affine transformation with the same random state corresponding to those applied with the raw image x. These augmentations operate on `PyTorch` tensors and are publicly-available at https://github.com/PhoenixDL/rising.git

8. Pre-trained cluster assignment maps for different K

We show in Fig. IV-2 the pre-trained cluster assignment for different number of clusters K. One can see that a small K learns a collapsed cluster assignment, which leads to a weak segmentation performance (see Table 6.3). This is probably because a small cluster number reduces the capacity to capture the structure information of such images. With the increase of K, the cluster maps become more balanced and gradually reflect the cardiac structures of the image. However, when taking a large cluster numbers $K = 60$, the resulted clusters over-segment the images, leading to fractured anatomical structures. In this case, it can decrease the downstream segmentation tasks.

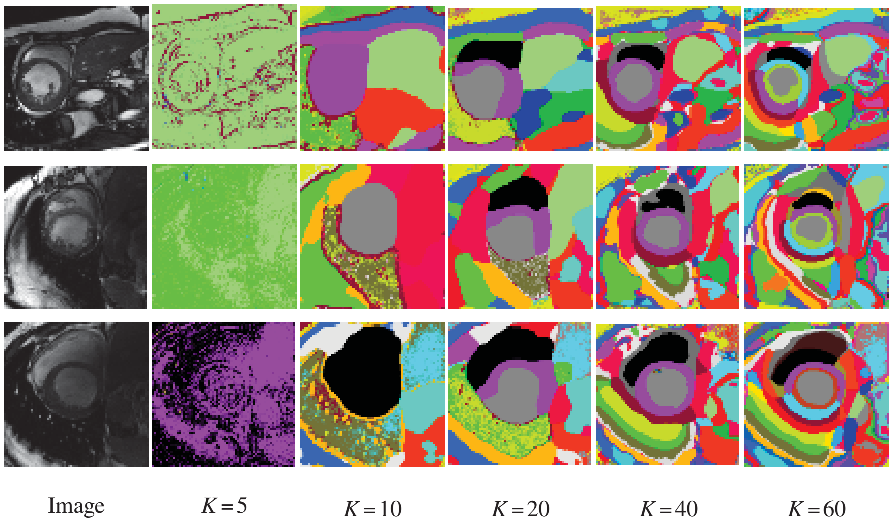


Figure-A IV-2 Pre-trained cluster assignment with respect to different K

9. Impact of batch size on pre-training

As contrastive-based pre-training often requires a large batch size, which is hard to satisfy for dense prediction tasks, such as segmentation. Our last ablation study investigates the performance stability given relatively small batch size $\mathcal{B}$. Table IV-1 lists the 3D test DSC for ACDC dataset with reduced batch size for contrastive-based and one of our best performing variant. It can be seen that with reduced batch size, segmentation performances reduces for both methods. However, our proposed method still outperforms contrastive-based approach for almost all cases given a very small batch.

Table-A IV-1 Impact of batch size $\mathcal{B}$

$\mathcal{B}$	**ACDC-LV**			**ACDC-RV**			**ACDC-Myo**		
	1 scan	**2 scans**	**4 scans**	**1 scan**	**2 scans**	**4 scans**	**1 scans**	**2 scans**	**4 scans**
	Contrast (*Enc*+*Dec*)								
6	72.16	86.02	87.37	63.30	69.25	72.25	61.51	73.57	76.51
12	75.52	84.23	88.31	63.26	71.41	73.89	65.88	76.26	78.77
18	77.98	85.97	88.42	66.47	72.82	76.69	64.96	76.98	78.76
	Ours (*MI*+*CC*)								
6	81.61	85.76	88.21	67.39	67.04	66.08	71.18	77.41	80.20
12	81.46	87.89	88.72	68.15	76.33	74.96	74.84	78.54	82.58
18	84.04	88.52	89.31	76.86	79.13	75.92	76.93	79.59	81.97

10. Visual results for segmentation

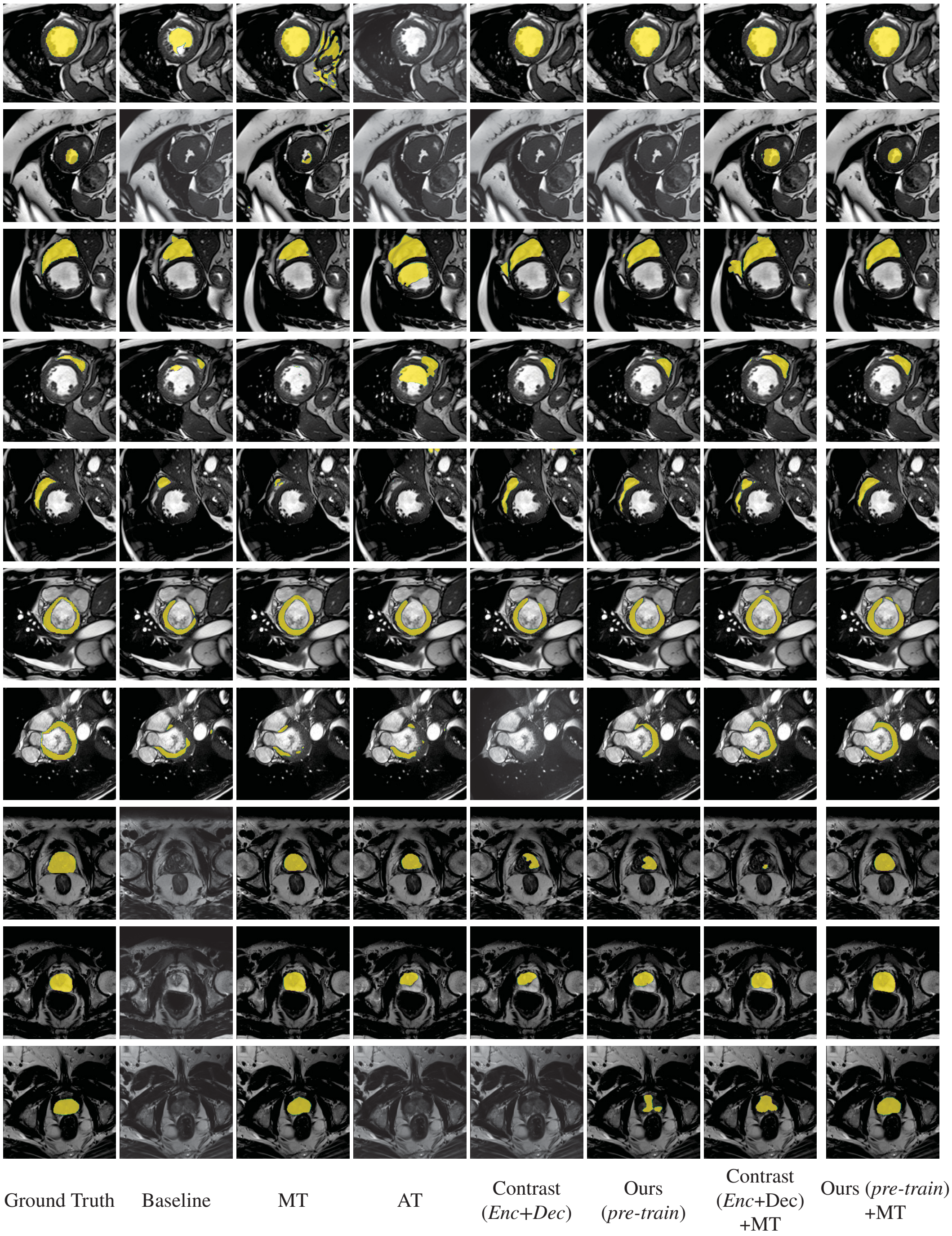


Figure-A IV-3 Visual comparison of tested methods on test images. Rows 1–2: **LV**; Rows 3–5: **RV**; Rows 6–7: **Myo**; Row 8-10: **Promise12**.

APPENDIX V

CODE AVAILABILITY

1. Presentation

We publish most of our research code to support the reproducibility of our research works and promote the open-source spirit in our communities. This appendix presents the code link for selected projects.

2. Code link

The code can be found with the following links, based on which, researchers can just modify small components of the code to conduct fair comparisons for various semi- /weakly- supervised algorithms on different medical segmentation tasks.

- Discretely-constrained deep network for weakly supervised segmentation:
  https://github.com/jizongFox/DGA1033
- Deep co-training for semi-supervised image segmentation:
  https://github.com/jizongFox/Deep-Co-Training-for-Semi-Supervised-Image-Segmentation
- Boosting Semi-supervised Image Segmentation with Global and Local Mutual Information Regularization:
  https://github.com/jizongFox/MI-based-Regularized-Semi-supervised-Segmentation
- Self-Paced Contrastive Learning for Semi-supervised Medical Image Segmentation with Meta-labels:
  https://github.com/jizongFox/Self-paced-Contrastive-Learning
- Information-based Deep Clustering: An Experimental Study:
  https://github.com/jizongFox/DeepClusteringProject

BIBLIOGRAPHY

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P. & Süsstrunk, S. (2012). SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11), 2274–2282.

Agarap, A. F. (2018). Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*.

Ahn, E., Feng, D. & Kim, J. (2021). A Spatial Guided Self-supervised Clustering Network for Medical Image Segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 379–388.

Ahsan, U., Madhok, R. & Essa, I. (2019). Video Jigsaw: Unsupervised Learning of Spatiotemporal Context for Video Action Recognition. *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 179–189.

Al-Ayyoub, M., Hmeidi, I. & Rababah, H. (2013). Detecting Hand Bone Fractures in X-Ray Images. *J. Multim. Process. Technol.*, 4(3), 155–168.

Alemi, A. A., Fischer, I., Dillon, J. V. & Murphy, K. (2016). Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*.

Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Landman, B. A., Litjens, G., Menze, B., Ronneberger, O., Summers, R. M., van Ginneken, B. et al. (2021). The Medical Segmentation Decathlon. *arXiv preprint arXiv:2106.05735*.

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. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 253–260.

Bai, W., Chen, C., Tarroni, G., Duan, J., Guitton, F., Petersen, S. E., Guo, Y., Matthews, P. M. & Rueckert, D. (2019). Self-supervised learning for cardiac MR image segmentation by anatomical position prediction. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 541–549.

Balakrishnan, G., Zhao, A., Sabuncu, M. R., Guttag, J. & Dalca, A. V. (2019). Voxelmorph: a learning framework for deformable medical image registration. *IEEE transactions on medical imaging*, 38(8), 1788–1800.

Banfield, J. D. & Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 803–821.

Bateson, M., Kervadec, H., Dolz, J., Lombaert, H. & Ayed, I. B. (2019). Constrained domain adaptation for segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 326–334.

Bateson, M., Kervadec, H., Dolz, J., Lombaert, H. & Ayed, I. B. (2020). Source-relaxed domain adaptation for image segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 490–499.

Baumgartner, C. F., Koch, L. M., Pollefeys, M. & Konukoglu, E. (2017). An exploration of 2D and 3D deep learning techniques for cardiac MR image segmentation. *International Workshop on Statistical Atlases and Computational Models of the Heart*, pp. 111–119.

Baur, C., Albarqouni, S. & Navab, N. (2017). Semi-supervised deep learning for fully convolutional networks. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 311–319.

Bearman, A., Russakovsky, O., Ferrari, V. & Fei-Fei, L. (2016). What's the point: Semantic segmentation with point supervision. *European Conference on Computer Vision*, pp. 549–565.

Belghazi, M. I., Baratin, A., Rajeswar, S., Ozair, S., Bengio, Y., Courville, A. & Hjelm, R. D. (2018). MINE: mutual information neural estimation. *arXiv preprint arXiv:1801.04062*.

Bengio, Y., Louradour, J., Collobert, R. & Weston, J. (2009). Curriculum learning. *Proceedings of the 26th annual international conference on machine learning*, pp. 41–48.

Bengio, Y., Courville, A. & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8), 1798–1828.

Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.-A., Cetin, I., Lekadir, K., Camara, O., Ballester, M. A. G. et al. (2018). Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved? *IEEE transactions on medical imaging*, 37(11), 2514–2525.

Bertels, J., Eelbode, T., Berman, M., Vandermeulen, D., Maes, F., Bisschops, R. & Blaschko, M. B. (2019). Optimizing the dice score and jaccard index for medical image segmentation: Theory and practice. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 92–100.

Blum, A. & Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. *Proceedings of the eleventh annual conference on 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. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 810–818.

Bose, A. J., Ling, H. & Cao, Y. (2018). Adversarial contrastive estimation. *arXiv preprint arXiv:1805.03642*.

Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J. et al. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1), 1–122.

Boykov, Y., Veksler, O. & Zabih, R. (2001). Fast approximate energy minimization via graph cuts. *IEEE Transactions on pattern analysis and machine intelligence*, 23(11), 1222–1239.

Bridle, J. S., Heading, A. J. & MacKay, D. J. (1992). Unsupervised Classifiers, Mutual Information and'Phantom Targets. *Advances in neural information processing systems*, pp. 1096–1101.

Can, Y. B., Chaitanya, K., Mustafa, B., Koch, L. M., Konukoglu, E. & Baumgartner, C. F. (2018). Learning to segment medical images with scribble-supervision alone. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support* (pp. 236–244). Springer.

Caron, M., Bojanowski, P., Joulin, A. & Douze, M. (2018). Deep clustering for unsupervised learning of visual features. *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 132–149.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P. & Joulin, A. (2020). Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint arXiv:2006.09882*.

Chaitanya, K., Karani, N., Baumgartner, C. F., Becker, A., Donati, O. & Konukoglu, E. (2019). Semi-supervised and task-driven data augmentation. *International Conference on Information Processing in Medical Imaging*, pp. 29–41.

Chaitanya, K., Erdil, E., Karani, N. & Konukoglu, E. (2020). Contrastive learning of global and local features for medical image segmentation with limited annotations. *arXiv preprint arXiv:2006.10511*, 33, 12546–12558.

Chapelle, O., Schlkopf, B. & Zien, A. (2010). *Semi-Supervised Learning* (ed. 1st). The MIT Press.

Chen, L., Bentley, P., Mori, K., Misawa, K., Fujiwara, M. & Rueckert, D. (2019a). Self-supervised learning for medical image analysis using image context restoration. *Medical image analysis*, 58, 101539.

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, T., Kornblith, S., Norouzi, M. & Hinton, G. (2020a). A Simple Framework for Contrastive Learning of Visual Representations. *Proceedings of the 37th International Conference on Machine Learning*, pp. 1597–1607.

Chen, T., Kornblith, S., Swersky, K., Norouzi, M. & Hinton, G. (2020b). Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*.

Chen, X., Udupa, J. K., Bagci, U., Zhuge, Y. & Yao, J. (2012). Medical image segmentation by combining graph cuts and oriented active appearance models. *IEEE transactions on image processing*, 21(4), 2035–2046.

Chen, X., Fan, H., Girshick, R. & He, K. (2020c). Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*.

Chen, X., Williams, B. M., Vallabhaneni, S. R., Czanner, G., Williams, R. & Zheng, Y. (2019b). Learning active contour models for medical image segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11632–11640.

Cheng, F., Chen, C., Wang, Y., Shi, H., Cao, Y., Tu, D., Zhang, C. & Xu, Y. (2020). Learning directional feature maps for cardiac MRI segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 108–117.

Cheng, R., Lay, N. S., Roth, H. R., Turkbey, B., Jin, D., Gandler, W., McCreedy, E. S., Pohida, T. J., Pinto, P. A., Choyke, P. L. et al. (2019). Fully automated prostate whole gland and central gland segmentation on MRI using holistically nested networks with short connections. *Journal of Medical Imaging*, 6(2), 024007.

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.

Cho, J. H., Mall, U., Bala, K. & Hariharan, B. (2021). PiCIE: Unsupervised Semantic Segmentation using Invariance and Equivariance in Clustering. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16794–16804.

Chou, E., Tramèr, F., Pellegrino, G. & Boneh, D. (2018). Sentinet: Detecting physical attacks against deep learning systems. *arXiv preprint arXiv:1812.00292*.

Chu, P. C. & Beasley, J. E. (1998). A genetic algorithm for the multidimensional knapsack problem. *Journal of heuristics*, 4(1), 63–86.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T. & Ronneberger, O. (2016). 3D U-Net: learning dense volumetric segmentation from sparse annotation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 424–432.

Clough, J. R., Byrne, N., Oksuz, I., Zimmer, V. A., Schnabel, J. A. & King, A. P. (2019). A topological loss function for deep-learning based image segmentation using persistent homology. *arXiv preprint arXiv:1910.01877*.

Cooper, D. & Freeman, J. (1970). On the Asymptotic Improvement in the Outcome of Supervised Learning Provided by Additional Nonsupervised Learning. *IEEE Transactions on Computers*, 19(11), 1055–1063. doi: 10.1109/T-C.1970.222832.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S. & Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213–3223.

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. *International Conference on Information Processing in Medical Imaging*, pp. 554–565.

Cunningham, W. H. (1985). On submodular function minimization. *Combinatorica*, 5(3), 185–192.

Dai, J., He, K. & Sun, J. (2015). Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1635–1643.

Decourt, C. & Duong, L. (2020). Semi-supervised generative adversarial networks for the segmentation of the left ventricle in pediatric MRI. *Computers in Biology and Medicine*, 123, 103884.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1), 1–38.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 248–255.

Derraz, F., Beladgham, M. & Khelif, M. (2004). Application of active contour models in medical image segmentation. *International Conference on Information Technology: Coding and Computing, 2004. Proceedings. ITCC 2004.*, 2, 675–681.

Devries, T. & Taylor, G. W. (2017). Improved Regularization of Convolutional Neural Networks with Cutout. *CoRR*, abs/1708.04552. Retrieved from: http://arxiv.org/abs/1708.04552.

Doersch, C. & Zisserman, A. (2017). Multi-task self-supervised visual learning. *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2051–2060.

Dolz, J., Ayed, I. B. & Desrosiers, C. (2017a). Unbiased shape compactness for segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 755–763.

Dolz, J., Ben Ayed, I. & Desrosiers, C. (2017b). DOPE: distributed optimization for pairwise energies. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6779–6788.

Dolz, J., Desrosiers, C. & Ben Ayed, I. (2018a). 3D fully convolutional networks for subcortical segmentation in MRI: A large-scale study. *NeuroImage*, 170, 456–470.

Dolz, J., Gopinath, K., Yuan, J., Lombaert, H., Desrosiers, C. & Ayed, I. B. (2018b). HyperDense-Net: A hyper-densely connected CNN for multi-modal image segmentation. *IEEE transactions on medical imaging*.

Donahue, J., Krähenbühl, P. & Darrell, T. (2016). Adversarial feature learning. *arXiv preprint arXiv:1605.09782*.

Dosovitskiy, A., Springenberg, J. T., Riedmiller, M. & Brox, T. (2014). Discriminative Unsupervised Feature Learning with Convolutional Neural Networks. *Advances in Neural Information Processing Systems*, 27.

Du, S., Lee, J., Li, H., Wang, L. & Zhai, X. (2019). Gradient Descent Finds Global Minima of Deep Neural Networks. *Proceedings of the 36th International Conference on Machine Learning*, 97, 1675–1685.

Duane, F., Aznar, M. C., Bartlett, F., Cutter, D. J., Darby, S. C., Jagsi, R., Lorenzen, E. L., McArdle, O., McGale, P., Myerson, S. et al. (2017). A cardiac contouring atlas for radiotherapy. *Radiotherapy and Oncology*, 122(3), 416–422.

El Jurdi, R., Petitjean, C., Honeine, P., Cheplygina, V. & Abdallah, F. (2021). High-level prior-based loss functions for medical image segmentation: A survey. *Computer Vision and Image Understanding*, 210, 103248.

Erhan, D., Bengio, Y., Courville, A., Manzagol, P.-A., Vincent, P. & Bengio, S. (2010). Why Does Unsupervised Pre-Training Help Deep Learning? *J. Mach. Learn. Res.*, 11, 625–660.

Fang, T., Liang, Z., Shao, X., Dong, Z. & Li, J. (2021). Self-supervised Multi-view Clustering for Unsupervised Image Segmentation. *International Conference on Artificial Neural Networks*, pp. 113–125.

Feng, Z., Xu, C. & Tao, D. (2019). Self-supervised representation learning by rotation feature decoupling. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10364–10374.

Freiman, M., Joskowicz, L. & Sosna, J. (2009). A variational method for vessels segmentation: algorithm and application to liver vessels visualization. *Medical Imaging 2009: Visualization, Image-Guided Procedures, and Modeling*, 7261, 72610H.

Ganaye, P.-A., Sdika, M., Triggs, B. & Benoit-Cattin, H. (2019). Removing segmentation inconsistencies with semi-supervised non-adjacency constraint. *Medical image analysis*, 58, 101551.

Ghasedi Dizaji, K., Herandi, A., Deng, C., Cai, W. & Huang, H. (2017). Deep clustering via joint convolutional autoencoder embedding and relative entropy minimization. *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5736–5745.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. & Bengio, Y. (2014). Generative Adversarial Nets. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D. & Weinberger, K. Q. (Eds.), *Advances in Neural Information Processing Systems 27* (pp. 2672–2680). Curran Associates, Inc.

Goodfellow, I., Shlens, J. & Szegedy, C. (2015). Explaining and Harnessing Adversarial Examples. *International Conference on Learning Representations*, pp. 1.

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

Grandvalet, Y. & Bengio, Y. (2005). Semi-supervised learning by entropy minimization. *Advances in neural information processing systems*, 367, 529–536.

Grandvalet, Y. & Bengio, Y. (2006). Entropy Regularization. In Chapelle, O., Schölkopf, B. & Zien, A. (Eds.), *Semi-Supervised Learning* (pp. 151–168). MIT Press.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P. H., Buchatskaya, E., Doersch, C., Pires, B. A., Guo, Z. D., Azar, M. G. et al. (2020). Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*.

Gupta, S., Hoffman, J. & Malik, J. (2016). Cross modal distillation for supervision transfer. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2827–2836.

Hadji, I. & Wildes, R. P. (2018). What do we understand about convolutional networks? *arXiv preprint arXiv:1803.08834*.

Han, Y., Feng, X.-C. & Baciu, G. (2013). Variational and PCA based natural image segmentation. *Pattern Recognition*, 46(7), 1971–1984.

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. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 562–571.

He, K., Zhang, X., Ren, S. & Sun, J. (2015). Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, (ICCV '15), 1026–1034. doi: 10.1109/ICCV.2015.123.

He, K., Zhang, X., Ren, S. & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

He, K., Fan, H., Wu, Y., Xie, S. & Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738.

He, K., Lian, C., Adeli, E., Huo, J., Gao, Y., Zhang, B., Zhang, J. & Shen, D. (2021). MetricUNet: Synergistic image-and voxel-level learning for precise prostate segmentation via online sampling. *Medical image analysis*, 71, 102039.

Hjelm, R. D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A. & Bengio, Y. (2018). Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*.

Hoffer, E. & Ailon, N. (2015). Deep metric learning using triplet network. *International workshop on similarity-based pattern recognition*, pp. 84–92.

Hosseinzadeh Taher, M. R., Haghighi, F., Feng, R., Gotway, M. B. & Liang, J. (2021). A Systematic Benchmarking Analysis of Transfer Learning for Medical Image Analysis. In *Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health* (pp. 3–13). Springer.

Hou, Q., Cheng, M.-M., Hu, X., Borji, A., Tu, Z. & Torr, P. (2017). Deeply supervised salient object detection with short connections. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5300–5309.

Hounsfield, G. N. (1980). Computed medical imaging. *Medical physics*, 7(4), 283–290.

Hsu, C.-C., Hsu, K.-J., Tsai, C.-C., Lin, Y.-Y. & Chuang, Y.-Y. (2019). Weakly supervised instance segmentation using the bounding box tightness prior. *Advances in Neural Information Processing Systems*, 32, 6586–6597.

Hu, W., Miyato, T., Tokui, S., Matsumoto, E. & Sugiyama, M. (2017a). Learning discrete representations via information maximizing self-augmented training. *International conference on machine learning*, pp. 1558–1567.

Hu, W., Miyato, T., Tokui, S., Matsumoto, E. & Sugiyama, M. (2017b). Learning Discrete Representations via Information Maximizing Self-Augmented Training. *ICML*.

Hu, X., Zeng, D., Xu, X. & Shi, Y. (2021). Semi-supervised contrastive learning for label-efficient medical image segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 481–490.

Huang, H., Zheng, H., Lin, L., Cai, M., Hu, H., Zhang, Q., Chen, Q., Iwamoto, Y., Han, X., Chena, Y.-W. et al. (2021). Medical Image Segmentation with Deep Atlas Prior. *IEEE Transactions on Medical Imaging*.

Huang, R., Zheng, Y., Hu, Z., Zhang, S. & Li, H. (2020). Multi-organ segmentation via co-training weight-averaged models from few-organ datasets. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 146–155.

Huh, M., Agrawal, P. & Efros, A. A. (2016). What makes ImageNet good for transfer learning? *arXiv preprint arXiv:1608.08614*.

Hung, W.-C., Tsai, Y.-H., Liou, Y.-T., Lin, Y.-Y. & Yang, M.-H. (2018). Adversarial Learning for Semi-supervised Semantic Segmentation. *Proceedings of the British Machine Vision Conference (BMVC)*, pp. 1.

Jégou, S., Drozdzal, M., Vazquez, D., Romero, A. & Bengio, Y. (2017). The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 11–19.

Ji, X., Henriques, J. F. & Vedaldi, A. (2019). Invariant information clustering for unsupervised image classification and segmentation. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9865–9874.

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., Yu, S.-I., Lan, Z., Shan, S. & Hauptmann, A. (2014). Self-paced learning with diversity. *Advances in Neural Information Processing Systems*, pp. 2078–2086.

Jing, L. & Tian, Y. (2020). Self-supervised visual feature learning with deep neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*.

Journée, M., Nesterov, Y., Richtárik, P. & Sepulchre, R. (2010). Generalized power method for sparse principal component analysis. *Journal of Machine Learning Research*, 11(Feb), 517–553.

Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R. & Wu, A. Y. (2002). An efficient k-means clustering algorithm: Analysis and implementation. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 1(7), 881–892.

Karimi, D. & Salcudean, S. E. (2019). Reducing the hausdorff distance in medical image segmentation with convolutional neural networks. *IEEE Transactions on medical imaging*, 39(2), 499–513.

Karnyaczki, S. & Desrosiers, C. (2015). A sparse coding method for semi-supervised segmentation with multi-class histogram constraints. *2015 IEEE International Conference on Image Processing (ICIP)*, pp. 3215–3219.

Kervadec, H., Dolz, J., Tang, M., Granger, E., Boykov, Y. & Ayed, I. B. (2018). Size-constraint loss for weakly supervised CNN segmentation.

Kervadec, H., Bouchtiba, J., Desrosiers, C., Granger, E., Dolz, J. & Ayed, I. B. (2019a). Boundary loss for highly unbalanced segmentation. *International conference on medical imaging with deep learning*, pp. 285–296.

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

Kervadec, H., Dolz, J., Wang, S., Granger, E. & Ayed, I. B. (2020). Bounding boxes for weakly supervised segmentation: Global constraints get close to full supervision. *Medical Imaging with Deep Learning*, pp. 365–381.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C. & Krishnan, D. (2020). Supervised contrastive learning. *arXiv preprint arXiv:2004.11362*.

Kim, D., Cho, D., Yoo, D. & Kweon, I. S. (2018). Learning image representations by completing damaged jigsaw puzzles. *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 793–802.

Kim, M., Tack, J. & Hwang, S. J. (2020). Adversarial self-supervised contrastive learning. *arXiv preprint arXiv:2006.07589*.

Kingma, D. P. & Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kingma, D. P., Mohamed, S., Jimenez Rezende, D. & Welling, M. (2014). Semi-supervised Learning with Deep Generative Models. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D. & Weinberger, K. Q. (Eds.), *Advances in Neural Information Processing Systems 27* (pp. 3581–3589). Curran Associates, Inc.

Kolesnikov, A. & Lampert, C. H. (2016). Seed, expand and constrain: Three principles for weakly-supervised image segmentation. *European Conference on Computer Vision*, pp. 695–711.

Komodakis, N. & Gidaris, S. (2018). Unsupervised representation learning by predicting image rotations. *International Conference on Learning Representations (ICLR)*.

Kraus, O. Z., Ba, J. L. & Frey, B. J. (2016). Classifying and segmenting microscopy images with deep multiple instance learning. *Bioinformatics*, 32(12), i52–i59.

Krause, A., Perona, P. & Gomes, R. G. (2010). Discriminative Clustering by Regularized Information Maximization. In Lafferty, J. D., Williams, C. K. I., Shawe-Taylor, J., Zemel, R. S. & Culotta, A. (Eds.), *Advances in Neural Information Processing Systems 23* (pp. 775–783). Curran Associates, Inc. Retrieved from: http://papers.nips.cc/paper/4154-discriminative-clustering-by-regularized-information-maximization.pdf.

Krizhevsky, A., Hinton, G. et al. (2009). *Learning multiple layers of features from tiny images*.

Krizhevsky, A., Sutskever, I. & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 1097–1105.

Kuhn, H. W. & Yaw, B. (1955). The Hungarian method for the assignment problem. *Naval Res. Logist. Quart*, 83–97.

Kumar, M. P., Packer, B. & Koller, D. (2010). Self-Paced Learning for Latent Variable Models. *NIPS*, 1, 2.

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

Laine, S. & Aila, T. (2017). Temporal Ensembling for Semi-Supervised Learning. *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Landman, B., Xu, Z., Igelsias, J., Styner, M., Langerak, T. & Klein, A. (2015). 2015 miccai multi-atlas labeling beyond the cranial vault workshop and challenge.

Laude, E., Lange, J.-H., Schüpfer, J., Domokos, C., Leal-Taixé, L., Schmidt, F. R., Andres, B. & Cremers, D. (2018). Discrete-continuous ADMM for transductive inference in higher-order MRFs. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1614–1624.

Lauterbur, P. C. (1973). Image formation by induced local interactions: examples employing nuclear magnetic resonance. *nature*, 242(5394), 190–191.

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

LeCun, Y., Bengio, Y. & Hinton, G. (2015). Deep learning. *nature*, 521(7553), 436.

Lee, D.-H. (2013). Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. *Workshop on Challenges in Representation Learning, ICML*, 3, 2.

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.

Lee, H.-Y., Huang, J.-B., Singh, M. & Yang, M.-H. (2017). Unsupervised representation learning by sorting sequences. *Proceedings of the IEEE International Conference on Computer Vision*, pp. 667–676.

Lerousseau, M., Vakalopoulou, M., Classe, M., Adam, J., Battistella, E., Carré, A., Estienne, T., Henry, T., Deutsch, E. & Paragios, N. (2020). Weakly supervised multiple instance learning histopathological tumor segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 470–479.

Levin, A., Viola, P. A. & Freund, Y. (2003). Unsupervised Improvement of Visual Detectors using Co-Training. *ICCV*, 1, 626–633.

Li, B. N., Chui, C. K., Chang, S. & Ong, S. H. (2011). Integrating spatial fuzzy clustering with level set methods for automated medical image segmentation. *Computers in biology and medicine*, 41(1), 1–10.

Li, F., Qiao, H. & Zhang, B. (2018a). Discriminatively boosted image clustering with fully convolutional auto-encoders. *Pattern Recognition*, 83, 161–173.

Li, K., Wang, S., Yu, L. & Heng, P.-A. (2020a). Dual-teacher: Integrating intra-domain and inter-domain teachers for annotation-efficient cardiac segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 418–427.

Li, S., Zhang, C. & He, X. (2020b). Shape-aware semi-supervised 3d semantic segmentation for medical images. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 552–561.

Li, X., Yu, L., Chen, H., Fu, C.-W. & Heng, P.-A. (2018b). Semi-supervised skin lesion segmentation via transformation consistent self-ensembling model. *arXiv preprint arXiv:1808.03887*.

Li, X., Yu, L., Chen, H., Fu, C.-W., Xing, L. & Heng, P.-A. (2020c). Transformation-consistent self-ensembling model for semisupervised medical image segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 32(2), 523–534.

Li, Z., Ko, B. & Choi, H.-J. (2019). Naive semi-supervised deep learning using pseudo-label. *Peer-to-peer networking and applications*, 12(5), 1358–1368.

Liao, R., Moyer, D., Golland, P. & Wells, W. M. (2020). Demi: Discriminative estimator of mutual information. *arXiv preprint arXiv:2010.01766*.

Lin, D., Dai, J., Jia, J., He, K. & Sun, J. (2016). Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3159–3167.

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.

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

Liu, B., Dolz, J., Galdran, A., Kobbi, R. & Ayed, I. B. (2021a). The hidden label-marginal biases of segmentation losses. *arXiv preprint arXiv:2104.08717*.

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

Liu, N. & Han, J. (2016). Dhsnet: Deep hierarchical saliency network for salient object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 678–686.

Liu, X., Zhang, F., Hou, Z., Mian, L., Wang, Z., Zhang, J. & Tang, J. (2021b). Self-supervised learning: Generative or contrastive. *IEEE Transactions on Knowledge and Data Engineering*.

Liu, Z., Zhu, Z., Zheng, S., Liu, Y., Zhou, J. & Zhao, Y. (2021c). Margin Preserving Self-paced Contrastive Learning Towards Domain Adaptation for Medical Image Segmentation. *arXiv preprint arXiv:2103.08454*.

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.

Luc, P., Couprie, C., Chintala, S. & Verbeek, J. (2016). Semantic Segmentation using Adversarial Networks. *NIPS Workshop on Adversarial Training*.

Luc, P., Neverova, N., Couprie, C., Verbeek, J. & LeCun, Y. (2017). Predicting deeper into the future of semantic segmentation. *IEEE International Conference on Computer Vision (ICCV)*, 1.

Luo, X., Chen, J., Song, T. & Wang, G. (2020). Semi-supervised medical image segmentation through dual-task consistency. *arXiv preprint arXiv:2009.04448*.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D. & Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Maeireizo, B., Litman, D. & Hwa, R. (2004). Co-training for predicting emotions with spoken dialogue data. *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, pp. 28.

Mahapatra, D. (2017). Semi-supervised learning and graph cuts for consensus based medical image segmentation. *Pattern recognition*, 63, 700–709.

Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I. & Frey, B. (2015). Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*.

Manohar, V., Povey, D. & Khudanpur, S. (2015). Semi-supervised maximum mutual information training of deep neural network acoustic models. *Sixteenth Annual Conference of the International Speech Communication Association*.

Marin, D., Tang, M., Ayed, I. B. & Boykov, Y. (2019). Beyond Gradient Descent for Regularized Segmentation Losses. *IEEE conference on Computer Vision and Pattern Recognition (CVPR)*.

McAllester, D. & Stratos, K. (2020). Formal limitations on the measurement of mutual information. *International Conference on Artificial Intelligence and Statistics*, pp. 875–884.

Medrano-Gracia, P., Cowan, B. R., Ambale-Venkatesh, B., Bluemke, D. A., Eng, J., Finn, J. P., Fonseca, C. G., Lima, J. A., Suinesiaputra, A. & Young, A. A. (2014). Left ventricular shape variation in asymptomatic populations: the multi-ethnic study of atherosclerosis. *Journal of Cardiovascular Magnetic Resonance*, 16(1), 56.

Miksik, O., Vineet, V., Pérez, P., Torr, P. & Sévigné, F. C. (2014). Distributed non-convex ADMM-inference in large-scale random fields. *British Machine Vision Conference (BMVC)*, 2(7).

Milletari, F., Navab, N. & Ahmadi, S.-A. (2016). V-net: Fully convolutional neural networks for volumetric medical image segmentation. *3D Vision (3DV), 2016 Fourth International Conference on*, pp. 565–571.

Min, S. & Chen, X. (2018). A Robust Deep Attention Network to Noisy Labels in Semi-supervised Biomedical Segmentation. *arXiv preprint arXiv:1807.11719*.

Mirikharaji, Z. & Hamarneh, G. (2018). Star shape prior in fully convolutional networks for skin lesion segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 737–745.

Misra, I. & Maaten, L. v. d. (2020). Self-supervised learning of pretext-invariant representations. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6707–6717.

Miyato, T., ichi Maeda, S., Koyama, M., Nakae, K. & Ishii, S. (2015). Distributional Smoothing by Virtual Adversarial Examples. *CoRR*, abs/1507.00677.

Miyato, T., Maeda, S.-i., Koyama, M. & Ishii, S. (2018). Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8), 1979–1993.

Morid, M. A., Borjali, A. & Del Fiol, G. (2020). A scoping review of transfer learning research on medical image analysis using ImageNet. *Computers in Biology and Medicine*, 104115.

Mukherjee, S., Asnani, H. & Kannan, S. (2020). CCMI: Classifier based conditional mutual information estimation. *Uncertainty in Artificial Intelligence*, pp. 1083–1093.

Navarro, F., Shit, S., Ezhov, I., Paetzold, J., Gafita, A., Peeken, J. C., Combs, S. E. & Menze, B. H. (2019). Shape-aware complementary-task learning for multi-organ segmentation. *International Workshop on Machine Learning in Medical Imaging*, pp. 620–627.

Neary, P. (2018). Automatic hyperparameter tuning in deep convolutional neural networks using asynchronous reinforcement learning. *2018 IEEE international conference on cognitive computing (ICCC)*, pp. 73–77.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B. & Ng, A. Y. (2011). Reading digits in natural images with unsupervised feature learning. *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*.

Ng, D., Lan, X., Yao, M. M.-S., Chan, W. P. & Feng, M. (2021). Federated learning: a collaborative effort to achieve better medical imaging models for individual sites that have small labelled datasets. *Quantitative Imaging in Medicine and Surgery*, 11(2), 852.

Nguyen, K.-C. T., Shi, D., Kaipatur, N. R., Lou, E. H., Major, P. W., Punithakumar, K. & Le, L. H. (2018). Graph cuts-based segmentation of alveolar bone in ultrasound imaging. *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 2049–2055.

Nguyen, Q. & Hein, M. (2017). The loss surface of deep and wide neural networks. *International conference on machine learning*, pp. 2603–2612.

Nigam, K. & Ghani, R. (2000). Understanding the behavior of co-training. *Proceedings of KDD-2000 workshop on text mining*, pp. 15–17.

Noh, H., Hong, S. & Han, B. (2015). Learning deconvolution network for semantic segmentation. *Proceedings of the IEEE international conference on computer vision*, pp. 1520–1528.

Noroozi, M. & Favaro, P. (2016). Unsupervised learning of visual representations by solving jigsaw puzzles. *European Conference on Computer Vision*, pp. 69–84.

Noroozi, M., Vinjimoor, A., Favaro, P. & Pirsiavash, H. (2018). Boosting self-supervised learning via knowledge transfer. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9359–9367.

Nowozin, S., Cseke, B. & Tomioka, R. (2016). f-gan: Training generative neural samplers using variational divergence minimization. *Advances in neural information processing systems*, pp. 271–279.

Odena, A. (2016). Semi-supervised learning with generative adversarial networks. *arXiv preprint arXiv:1606.01583*.

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.

Oliva, A. & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3), 145–175.

Oliver, A., Odena, A., Raffel, C. A., Cubuk, E. D. & Goodfellow, I. (2018). Realistic Evaluation of Deep Semi-Supervised Learning Algorithms. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N. & Garnett, R. (Eds.), *Advances in Neural Information Processing Systems 31* (pp. 3239–3250). Curran Associates, Inc.

Oord, A. v. d., Li, Y. & Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Ouyang, C., Biffi, C., Chen, C., Kart, T., Qiu, H. & Rueckert, D. (2020). Self-supervision with superpixels: Training few-shot medical image segmentation without annotation. *European Conference on Computer Vision*, pp. 762–780.

Ouyang, H., He, N., Tran, L. & Gray, A. (2013). Stochastic alternating direction method of multipliers. *International Conference on Machine Learning*, pp. 80–88.

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.

Paine, T. L., Khorrami, P., Han, W. & Huang, T. S. (2014). An Analysis of Unsupervised Pre-training in Light of Recent Advances. *arXiv preprint arXiv:1412.6597*.

Paninski, L. (2003). Estimation of Entropy and Mutual Information. *Neural Comput.*, 15(6), 1191–1253. doi: 10.1162/089976603321780272.

Papandreou, G., Chen, L.-C., Murphy, K. P. & Yuille, A. L. (2015). Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. *Proceedings of the IEEE international conference on computer vision*, pp. 1742–1750.

Paragios, N. (2002). A variational approach for the segmentation of the left ventricle in cardiac image analysis. *International Journal of Computer Vision*, 50(3), 345–362.

Paszke, A., Chaurasia, A., Kim, S. & Culurciello, E. (2016). Enet: A deep neural network architecture for real-time semantic segmentation. *arXiv preprint arXiv:1606.02147*.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L. & Lerer, A. (2017). Automatic differentiation in pytorch.

Patel, G. & Dolz, J. (2021). Weakly supervised segmentation with cross-modality equivariant constraints. *arXiv preprint arXiv:2104.02488*.

Pathak, D., Shelhamer, E., Long, J. & Darrell, T. (2014). Fully convolutional multi-class multiple instance learning. *arXiv preprint arXiv:1412.7144*.

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.

Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T. & Efros, A. A. (2016). Context encoders: Feature learning by inpainting. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2536–2544.

Pathak, D., Agrawal, P., Efros, A. A. & Darrell, T. (2017). Curiosity-driven exploration by self-supervised prediction. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 16–17.

Peng, J. & Wang, Y. (2021). Medical image segmentation with limited supervision: A review of deep network models. *IEEE Access*.

Peng, J., Desrosiers, C. & Pedersoli, M. (2019). Information based Deep Clustering: An experimental study. *arXiv preprint arXiv:1910.01665*.

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

Peng, J., Pedersoli, M. & Desrosiers, C. (2020b). Mutual information deep regularization for semi-supervised segmentation. *Proceedings of Machine Learning Research*, 1, 13.

Peng, J., Pedersoli, M. & Desrosiers, C. (2021a). Boosting Semi-supervised Image Segmentation with Global and Local Mutual Information Regularization. *arXiv preprint arXiv:2103.04813*.

Peng, J., Wang, P., Desrosiers, C. & Pedersoli, M. (2021b). Self-Paced Contrastive Learning for Semi-supervisedMedical Image Segmentation with Meta-labels. *arXiv preprint arXiv:2107.13741*.

Peng, X., Tang, Z., Yang, F., Feris, R. S. & Metaxas, D. (2018). Jointly optimize data augmentation and network training: Adversarial data augmentation in human pose estimation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2226–2234.

Penha, G. & Hauff, C. (2019). Curriculum Learning Strategies for IR: An Empirical Study on Conversation Response Ranking. *arXiv preprint arXiv:1912.08555*.

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.

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

Pham, D. L., Xu, C. & Prince, J. L. (2000). Current methods in medical image segmentation. *Annual review of biomedical engineering*, 2(1), 315–337.

Pihur, V., Datta, S. & Datta, S. (2007). Weighted rank aggregation of cluster validation measures: a monte carlo cross-entropy approach. *Bioinformatics*, 23(13), 1607–1615.

Pinheiro, P. O. & Collobert, R. (2015a). From image-level to pixel-level labeling with convolutional networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1713–1721.

Pinheiro, P. O. & Collobert, R. (2015b). Weakly supervised semantic segmentation with convolutional networks. *Proceedings of the IEEE international conference on computer vision*, 2(5), 6.

Platanios, E. A., Stretcu, O., Neubig, G., Poczos, B. & Mitchell, T. M. (2019). Competence-based curriculum learning for neural machine translation. *arXiv preprint arXiv:1903.09848*.

Pluim, J. P., Maintz, J. A. & Viergever, M. A. (2003). Mutual-information-based registration of medical images: a survey. *IEEE transactions on medical imaging*, 22(8), 986–1004.

Poole, B., Ozair, S., Oord, A. v. d., Alemi, A. A. & Tucker, G. (2019). On variational bounds of mutual information. *arXiv preprint arXiv:1905.06922*.

Prados, F., Ashburner, J., Blaiotta, C., Brosch, T., Carballido-Gamio, J., Cardoso, M. J., Conrad, B. N., Datta, E., Dávid, G., De Leener, B. et al. (2017). Spinal cord grey matter segmentation challenge. *Neuroimage*, 152, 312–329.

Qi, X., Liu, Z., Shi, J., Zhao, H. & Jia, J. (2016). Augmented feedback in semantic segmentation under image level supervision. *European Conference on Computer Vision*, pp. 90–105.

Qian, X., Wang, J., Guo, S. & Li, Q. (2013). An active contour model for medical image segmentation with application to brain CT image. *Medical physics*, 40(2), 021911.

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 (ECCV)*, pp. 135–152.

Qiu, J., Sun, K., Wang, T. & Gao, H. (2019). Observer-based fuzzy adaptive event-triggered control for pure-feedback nonlinear systems with prescribed performance. *IEEE Transactions on Fuzzy systems*, 27(11), 2152–2162.

Qu, H., Wu, P., Huang, Q., Yi, J., Riedlinger, G. M., De, S. & Metaxas, D. N. (2019). Weakly supervised deep nuclei segmentation using points annotation in histopathology images. *International Conference on Medical Imaging with Deep Learning*, pp. 390–400.

Radosavovic, I., Dollár, P., Girshick, R., Gkioxari, G. & He, K. (2018). Data distillation: Towards omni-supervised learning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4119–4128.

Rajchl, M., Lee, M. C., Oktay, O., Kamnitsas, K., Passerat-Palmbach, J., Bai, W., Damodaram, M., Rutherford, M. A., Hajnal, J. V., Kainz, B. et al. (2017). Deepcut: Object segmentation from bounding box annotations using convolutional neural networks. *IEEE transactions on medical imaging*, 36(2), 674–683.

Rasmus, A., Berglund, M., Honkala, M., Valpola, H. & Raiko, T. (2015). Semi-supervised Learning with Ladder Networks. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M. & Garnett, R. (Eds.), *Advances in Neural Information Processing Systems 28* (pp. 3546–3554). Curran Associates, Inc.

Ren, M., Triantafillou, E., Ravi, S., Snell, J., Swersky, K., Tenenbaum, J. B., Larochelle, H. & Zemel, R. S. (2018). Meta-learning for semi-supervised few-shot classification. *arXiv preprint arXiv:1803.00676*.

Rifai, S., Glorot, X., Bengio, Y. & Vincent, P. (2011). Adding noise to the input of a model trained with a regularized objective. *arXiv preprint arXiv:1104.3250*.

Ronneberger, O., Fischer, P. & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In Navab, N., Hornegger, J., Wells, W. M. & Frangi, A. F. (Eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III* (pp. 234–241). Cham: Springer International Publishing. doi: 10.1007/978-3-319-24574-4_28.

Röntgen, W. C. (1896). On a new kind of rays. *Science*, 3(59), 227–231.

Roth, H. R., Yang, D., Xu, Z., Wang, X. & Xu, D. (2021). Going to Extremes: Weakly Supervised Medical Image Segmentation. *Machine Learning and Knowledge Extraction*, 3(2), 507–524.

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

Saleh, F., Aliakbarian, M. S., Salzmann, M., Petersson, L., Gould, S. & Alvarez, J. M. (2016). Built-in foreground/background prior for weakly-supervised semantic segmentation. *European Conference on Computer Vision*, pp. 413–432.

Salehi, S. S. M., Erdogmus, D. & Gholipour, A. (2017). Tversky loss function for image segmentation using 3D fully convolutional deep networks. *International workshop on machine learning in medical imaging*, pp. 379–387.

Saxena, A. (2016). Convolutional neural networks: an illustration in TensorFlow. *XRDS: Crossroads, The ACM Magazine for Students*, 22(4), 56–58.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural networks*, 61, 85–117.

Schroff, F., Criminisi, A. & Zisserman, A. (2011). Harvesting Image Databases from the Web. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(4), 754–766. doi: 10.1109/TPAMI.2010.133.

Schumacher, M., Genz, A. & Heinrich, M. (2020). Weakly supervised pancreas segmentation based on class activation maps. *Medical Imaging 2020: Image Processing*, 11313, 1131314.

Schwarz, M., Schulz, H. & Behnke, S. (2015). RGB-D object recognition and pose estimation based on pre-trained convolutional neural network features. *Robotics and Automation (ICRA), 2015 IEEE International Conference on*, pp. 1329–1335.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D. et al. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *ICCV*, pp. 618–626.

Sermanet, P., Lynch, C., Chebotar, Y., Hsu, J., Jang, E., Schaal, S., Levine, S. & Brain, G. (2018). Time-contrastive networks: Self-supervised learning from video. *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1134–1141.

Shen, H., Wang, R., Zhang, J. & McKenna, S. J. (2017). Boundary-aware fully convolutional network for brain tumor segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 433–441.

Shi, J. & Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8), 888–905.

Shi, X. & Li, C. (2021). Convexity preserving level set for left ventricle segmentation. *Magnetic Resonance Imaging*, 78, 109–118.

Shimoda, W. & Yanai, K. (2016). Distinct class-specific saliency maps for weakly supervised semantic segmentation. *European Conference on Computer Vision*, pp. 218–234.

Simantiris, G. & Tziritas, G. (2020). Cardiac MRI segmentation with a dilated cnn incorporating domain-specific constraints. *IEEE Journal of Selected Topics in Signal Processing*, 14(6), 1235–1243.

Simpson, A. L., Antonelli, M., Bakas, S., Bilello, M., Farahani, K., van Ginneken, B., Kopp-Schneider, A., Landman, B. A., Litjens, G., Menze, B. et al. (2019). A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063*.

Singh, K. K. & Lee, Y. J. (2017). Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 3544–3553.

Siu, C. Y., Chan, H. L. & Ming Lui, R. L. (2020). Image Segmentation with Partial Convexity Shape Prior Using Discrete Conformality Structures. *SIAM Journal on Imaging Sciences*, 13(4), 2105–2139.

Snell, J., Swersky, K. & Zemel, R. S. (2017). Prototypical networks for few-shot learning. *arXiv preprint arXiv:1703.05175*.

Song, J. & Ermon, S. (2019). Understanding the limitations of variational mutual information estimators. *arXiv preprint arXiv:1910.06222*.

Soudry, D. & Carmon, Y. (2016). No bad local minima: Data independent training error guarantees for multilayer neural networks.

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

Springenberg, J. T. (2015). Unsupervised and semi-supervised learning with categorical generative adversarial networks. *arXiv preprint arXiv:1511.06390*.

Su, J., Vargas, D. V. & Sakurai, K. (2019). One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*.

Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S. & Cardoso, M. J. (2017). Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep learning in medical image analysis and multimodal learning for clinical decision support* (pp. 240–248). Springer.

Suetens, P. (2017). *Fundamentals of medical imaging*. Cambridge university press.

Sun, K., Mou, S., Qiu, J., Wang, T. & Gao, H. (2018). Adaptive fuzzy control for nontriangular structural stochastic switched nonlinear systems with full state constraints. *IEEE Transactions on Fuzzy systems*, 27(8), 1587–1601.

Sun, K., Shi, H., Zhang, Z. & Huang, Y. (2021). Ecs-net: Improving weakly supervised semantic segmentation by using connections between class activation maps. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7283–7292.

Suzuki, T., Sugiyama, M., Sese, J. & Kanamori, T. (2008). Approximating mutual information by maximum likelihood density ratio estimation. *New challenges for feature selection in data mining and knowledge discovery*, pp. 5–20.

Taghanaki, S. A., Zheng, Y., Zhou, S. K., Georgescu, B., Sharma, P., Xu, D., Comaniciu, D. & Hamarneh, G. (2019). Combo loss: Handling input and output imbalance in multi-organ segmentation. *Computerized Medical Imaging and Graphics*, 75, 24–33.

Taleb, A., Lippert, C., Klein, T. & Nabi, M. (2019). Multimodal self-supervised learning for medical image analysis. *arXiv preprint arXiv:1912.05396*.

Taleb, A., Lippert, C., Klein, T. & Nabi, M. (2021). Multimodal self-supervised learning for medical image analysis. *International Conference on Information Processing in Medical Imaging*, pp. 661–673.

Tang, M., Djelouah, A., Perazzi, F., Boykov, Y. & Schroers, C. (2018). Normalized cut loss for weakly-supervised cnn segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1818–1827.

Tarvainen, A. & Valpola, H. (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, pp. 1195–1204.

Thévenaz, P. & Unser, M. (2000). Optimization of mutual information for multiresolution image registration. *IEEE transactions on image processing*, 9(ARTICLE), 2083–2099.

Tian, Y., Krishnan, D. & Isola, P. (2019). Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*.

Tishby, N., Pereira, F. C. & Bialek, W. (2000). The information bottleneck method. *arXiv preprint physics/0004057*.

Valanarasu, J. M. J., Oza, P., Hacihaliloglu, I. & Patel, V. M. (2021). Medical transformer: Gated axial-attention for medical image segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 36–46.

Vejmelka, M. & Hlaváčková-Schindler, K. (2007). Mutual information estimation in higher dimensions: A speed-up of a k-nearest neighbor based estimator. *International conference on adaptive and natural computing algorithms*, pp. 790–797.

Veksler, O. (2008). Star shape prior for graph-cut image segmentation. *European Conference on Computer Vision*, pp. 454–467.

Verma, V., Kawaguchi, K., Lamb, A., Kannala, J., Bengio, Y. & Lopez-Paz, D. (2019). Interpolation consistency training for semi-supervised learning. *arXiv preprint arXiv:1903.03825*.

Vezhnevets, A. & Buhmann, J. M. (2010). Towards weakly supervised semantic segmentation by means of multiple instance and multitask learning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3249–3256.

Vincent, P., Larochelle, H., Bengio, Y. & Manzagol, P.-A. (2008). Extracting and Composing Robust Features with Denoising Autoencoders. *Proceedings of the 25th International Conference on Machine Learning*, pp. 1096–1103.

Viola, P. & Wells III, W. M. (1997). Alignment by maximization of mutual information. *International journal of computer vision*, 24(2), 137–154.

Vondrick, C., Patterson, D. & Ramanan, D. (2013). Efficiently Scaling Up Crowdsourced Video Annotation. *Int. J. Comput. Vision*, 101(1), 184–204. doi: 10.1007/s11263-012-0564-1.

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 Conference on Computer Vision and Pattern Recognition*, pp. 2517–2526.

Wan, X. (2009). Co-training for cross-lingual sentiment classification. *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-volume 1*, pp. 235–243.

Wang, F. & Liu, H. (2021). Understanding the behaviour of contrastive loss. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2495–2504.

Wang, J. & Xia, B. (2021). Bounding Box Tightness Prior for Weakly Supervised Image Segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 526–536.

Wang, L., Nie, D., Li, G., Puybareau, É., Dolz, J., Zhang, Q., Wang, F., Xia, J., Wu, Z., Chen, J. et al. (2019a). Benchmark on Automatic 6-month-old Infant Brain Segmentation Algorithms: The iSeg-2017 Challenge. *IEEE transactions on medical imaging*.

Wang, P., Peng, J., Pedersoli, M., Zhou, Y., Zhang, C. & Desrosiers, C. (2021). Self-paced and self-consistent co-training for semi-supervised image segmentation. *Medical Image Analysis*, 73, 102146.

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, X., Kihara, D., Luo, J. & Qi, G.-J. (2019b). Enaet: Self-trained ensemble autoencoding transformations for semi-supervised learning. *arXiv preprint arXiv:1911.09265*.

Wei, C., Xie, L., Ren, X., Xia, Y., Su, C., Liu, J., Tian, Q. & Yuille, A. L. (2019). Iterative Reorganization with Weak Spatial Constraints: Solving Arbitrary Jigsaw Puzzles for Unsupervised Representation Learning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1910–1919.

Wei, Y., Liang, X., Chen, Y., Jie, Z., Xiao, Y., Zhao, Y. & Yan, S. (2016). Learning to segment with image-level annotations. *Pattern Recognition*, 59, 234–244.

Wei, Y., Feng, J., Liang, X., Cheng, M.-M., Zhao, Y. & Yan, S. (2017). Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. *IEEE CVPR*, 1(2), 3.

Wei, Z., Shi, F., Song, H., Ji, W. & Han, G. (2020). Attentive boundary aware network for multi-scale skin lesion segmentation with adversarial training. *Multimedia Tools and Applications*, 79(37), 27115–27136.

Wolsey, L. A. & Nemhauser, G. L. (2014). *Integer and combinatorial optimization*. John Wiley & Sons.

Wu, J., Fan, H., Zhang, X., Lin, S. & Li, Z. (2021a). Semi-Supervised Semantic Segmentation via Entropy Minimization. *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6.

Wu, X., Dyer, E. & Neyshabur, B. (2021b). When Do Curricula Work? *International Conference on Learning Representations*. Retrieved from: https://openreview.net/forum?id=tW4QEInpni.

Xi, N. (2019). Semi-supervised Attentive Mutual-info Generative Adversarial Network for Brain Tumor Segmentation. *2019 International Conference on Image and Vision Computing New Zealand (IVCNZ)*, pp. 1–7.

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

Xiang, J., Li, Z., Wang, W., Xia, Q. & Zhang, S. (2021). Self-Ensembling Contrastive Learning for Semi-Supervised Medical Image Segmentation. *arXiv preprint arXiv:2105.12924*.

Xie, J., Girshick, R. & Farhadi, A. (2016). Unsupervised Deep Embedding for Clustering Analysis. *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, (ICML'16), 478–487. Retrieved from: http://dl.acm.org/citation.cfm?id=3045390.3045442.

Xiong, Y., Ren, M., Zeng, W. & Urtasun, R. (2021). Self-Supervised Representation Learning from Flow Equivariance. *arXiv preprint arXiv:2101.06553*.

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

Xu, Y., Gong, M., Fu, H., Tao, D., Zhang, K. & Batmanghelich, K. (2018). Multi-scale masked 3-D U-net for brain tumor segmentation. *International MICCAI Brainlesion Workshop*, pp. 222–233.

Xue, Y., Xu, T. & Huang, X. (2018). Adversarial learning with multi-scale loss for skin lesion segmentation. *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pp. 859–863.

Xue, Y., Tang, H., Qiao, Z., Gong, G., Yin, Y., Qian, Z., Huang, C., Fan, W. & Huang, X. (2020). Shape-aware organ segmentation by predicting signed distance maps. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07), 12565–12572.

Yang, D., Roth, H., Wang, X., Xu, Z., Myronenko, A. & Xu, D. (2020). Enhancing Foreground Boundaries for Medical Image Segmentation. *arXiv preprint arXiv:2005.14355*.

Yang, J., Parikh, D. & Batra, D. (2016). Joint Unsupervised Learning of Deep Representations and Image Clusters. *CoRR*, abs/1604.03628. Retrieved from: http://arxiv.org/abs/1604.03628.

Yeung, M., Sala, E., Schönlieb, C.-B. & Rundo, L. (2021). Unified Focal loss: Generalising Dice and cross entropy-based losses to handle class imbalanced medical image segmentation. *Computerized Medical Imaging and Graphics*, 102026.

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. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 605–613.

Yuan, X., He, P., Zhu, Q. & Li, X. (2019). Adversarial examples: Attacks and defenses for deep learning. *IEEE transactions on neural networks and learning systems*.

Yuan, Y., Chao, M. & Lo, Y.-C. (2017). Automatic skin lesion segmentation using deep fully convolutional networks with Jaccard distance. *IEEE Trans. Med. Imaging*, 36(9), 1876–1886.

Zeng, D., Wu, Y., Hu, X., Xu, X., Yuan, H., Huang, M., Zhuang, J., Hu, J. & Shi, Y. (2021). Positional Contrastive Learning for Volumetric Medical Image Segmentation. *arXiv preprint arXiv:2106.09157*.

Zhang, D., Meng, D., Zhao, L. & Han, J. (2017a). Bridging saliency detection to weakly supervised object detection based on self-paced curriculum learning. *arXiv preprint arXiv:1703.01290*.

Zhang, H., Cisse, M., Dauphin, Y. N. & Lopez-Paz, D. (2017b). mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.

Zhang, L., Qi, G.-J., Wang, L. & Luo, J. (2019). Aet vs. aed: Unsupervised representation learning by auto-encoding transformations rather than data. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2547–2555.

Zhang, M. & Desrosiers, C. (2018). High-quality image restoration using low-rank patch regularization and global structure sparsity. *IEEE Transactions on Image Processing*, 28(2), 868–879.

Zhang, M., Desrosiers, C. & Zhang, C. (2018). Atlas-based reconstruction of high performance brain MR data. *Pattern Recognition*, 76, 549–559.

Zhang, M., Dong, B. & Li, Q. (2020a). Deep active contour network for medical image segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 321–331.

Zhang, P., Zhong, Y. & Li, X. (2020b). ACCL: Adversarial constrained-CNN loss for weakly supervised medical image segmentation. *arXiv preprint arXiv:2005.00328*.

Zhang, R., Isola, P. & Efros, A. A. (2016). Colorful image colorization. *European conference on computer vision*, pp. 649–666.

Zhang, X., Smith, N. & Webb, A. (2008). Medical imaging. In *Biomedical Information Technology* (pp. 3–27). Elsevier.

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

Zhao, A., Balakrishnan, G., Durand, F., Guttag, J. V. & Dalca, A. V. (2019a). 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, H., Shi, J., Qi, X., Wang, X. & Jia, J. (2017). Pyramid scene parsing network. *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 2881–2890.

Zhao, J., Lu, D., Ma, K., Zhang, Y. & Zheng, Y. (2020a). Deep Image Clustering with Category-Style Representation. *European Conference on Computer Vision*, pp. 54–70.

Zhao, R., Qian, B., Zhang, X., Li, Y., Wei, R., Liu, Y. & Pan, Y. (2020b). Rethinking Dice Loss for Medical Image Segmentation. *2020 IEEE International Conference on Data Mining (ICDM)*, pp. 851–860.

Zhao, S., Wang, Y., Yang, Z. & Cai, D. (2019b). Region Mutual Information Loss for Semantic Segmentation. *Advances in Neural Information Processing Systems*, pp. 11115–11125.

Zhao, X., Vemulapalli, R., Mansfield, P., Gong, B., Green, B., Shapira, L. & Wu, Y. (2020c). Contrastive Learning for Label-Efficient Semantic Segmentation. *arXiv preprint arXiv:2012.06985*.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A. & Torralba, A. (2016). Learning deep features for discriminative localization. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2921–2929.

Zhou, Y., Wang, Y., Tang, P., Shen, W., Fishman, E. K. & Yuille, A. L. (2018a). Semi-supervised multi-organ segmentation via multi-planar co-training. *arXiv preprint arXiv:1804.02586*.

Zhou, Y., Li, Z., Bai, S., Wang, C., Chen, X., Han, M., Fishman, E. & Yuille, A. (2019a). Prior-aware Neural Network for Partially-Supervised Multi-Organ Segmentation. *arXiv preprint arXiv:1904.06346*.

Zhou, Y., Wang, Y., Tang, P., Bai, S., Shen, W., Fishman, E. & Yuille, A. (2019b). Semi-supervised 3D abdominal multi-organ segmentation via deep multi-planar co-training. *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 121–140.

Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N. & Liang, J. (2018b). Unet++: A nested u-net architecture for medical image segmentation. In *Deep learning in medical image analysis and multimodal learning for clinical decision support* (pp. 3–11). Springer.

Zhou, Z., Sodha, V., Pang, J., Gotway, M. B. & Liang, J. (2021). Models genesis. *Medical image analysis*, 67, 101840.

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.

Zhuang, X. & Shen, J. (2016). Multi-scale patch and multi-modality atlases for whole heart segmentation of MRI. *Medical image analysis*, 31, 77–87.

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, K. H., Warfield, S. K., Bharatha, A., Tempany, C. M., Kaus, M. R., Haker, S. J., Wells III, W. M., Jolesz, F. A. & Kikinis, R. (2004). Statistical validation of image segmentation quality based on a spatial overlap index1: scientific reports. *Academic radiology*, 11(2), 178–189.