Journal of Machine Learning for Biomedical Imaging. 2021:009. pp 1-29 Submitted 10/2020; Published 5/2021
Special Issue: Medical Imaging with Deep Learning (MIDL) 2020

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

Jizong Peng jizong.peng.1@etsmtl.net
Department of Software and IT Engineering, ETS Montreal, Canada
Marco Pedersoli marco.pedersoli@etsmtl.ca
Department of Systems Engineering, ETS Montreal, Canada
Christian Desrosiers christian.desrosiers@etsmtl.ca
Department of Software and IT Engineering, ETS Montreal, Canada

Abstract

The scarcity of labeled data often impedes the application of deep learning to the segmentation of medical images. Semi-supervised learning seeks to overcome this limitation by exploiting unlabeled examples in the learning process. 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. In this method, we maximize 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 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 provide an accuracy near to full supervision while training with very few annotated images

Keywords: Semantic segmentation, Semi-supervised learning, Deep clustering, Mutual information, Convolutional neural network

1. 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

Guest Editors: Marleen de Bruijne, Tal Arbel, Ismail Ben Ayed, Hervé Lombaert
https://www.melba-journal.org/article/23700.

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 and Cohen-Adad, 2018; Perone et al., 2019; Bortsova et al., 2019; Li et al., 2018; 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 and Cohen-Adad, 2018; Bortsova et al., 2019; Li et al., 2018). 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 et al., 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 et al., 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 et al., 2020), 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 et al., 2019; Ji et al., 2018).

An alternative approach to unsupervised representation learning, based on a discrete representation, is clustering (Ji et al., 2018; Caron et al., 2018; Peng et al., 2019a). 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 et al., 1992; Krause et al., 2010; Hu et al., 2017; Ji et al., 2018). 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. (2018) 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:

$$I(X;Y) \;=\; D_{\mathrm{KL}}\big(p(X,Y)\,||\,p(X)\,p(Y)\big). \tag{1}$$

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

$$\begin{aligned} I(X;Y) \;&= H(Y) \,-\, H(Y|X) \\ &= \mathbb{E}_Y\left[\log \mathbb{E}_X[\,p(Y|X)\,]\right] \,-\, \mathbb{E}_{X,Y}[\log p(Y|X)\,]. \end{aligned} \tag{2}$$

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. (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., 2017; Zhao et al., 2019b).

Recently, Peng et al. (2020) 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.

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., 2020) 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 3, 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 4. Results, reported in Section 5, show our method to significantly outperform compared approaches and yield performance near to full supervision when trained with only 5% of labeled examples.

2. Related works

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 and Chen, 2018), adversarial learning (Souly et al., 2017; Zhang et al., 2017), entropy minimization (Vu et al., 2019), co-training (Peng et al., 2019b; Zhou et al., 2019), temporal ensembling (Perone and 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 and 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 and 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.

Unsupervised representation learning Important efforts have also been invested towards learning robust representations from unlabeled data. In self-supervised learning (Noroozi and Favaro, 2016; Kim et al., 2018; Noroozi et al., 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 et al. (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 et al., 2019; Wang et al., 2019) and converting a grey-scale image to RGB (Zhang et al., 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. (2020) 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., 2017), 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., 2019), 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., 2018), instead maximizes the MI between cluster assignments of a sample and its transformed versions. Recently, Peng et al. (2020) 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.

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 et al., 2008; Vejmelka and 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 and Stratos, 2020), such as using f-divergence representation (Nowozin et al., 2016), Jensen–Shannon (JS) divergence based optimization (Hjelm et al., 2018; Zhao et al., 2020), and clipping output with a prefixed range (Song and Ermon, 2019). Contrastive-based methods have been shown to underestimate MI (Hjelm et al., 2018; McAllester and Stratos, 2020) and require a large number of negative examples (Tian et al., 2019). As alternative to MINE, discriminator-based MI estimation (Liao et al., 2020; Mukherjee et al., 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. 5.5, 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.

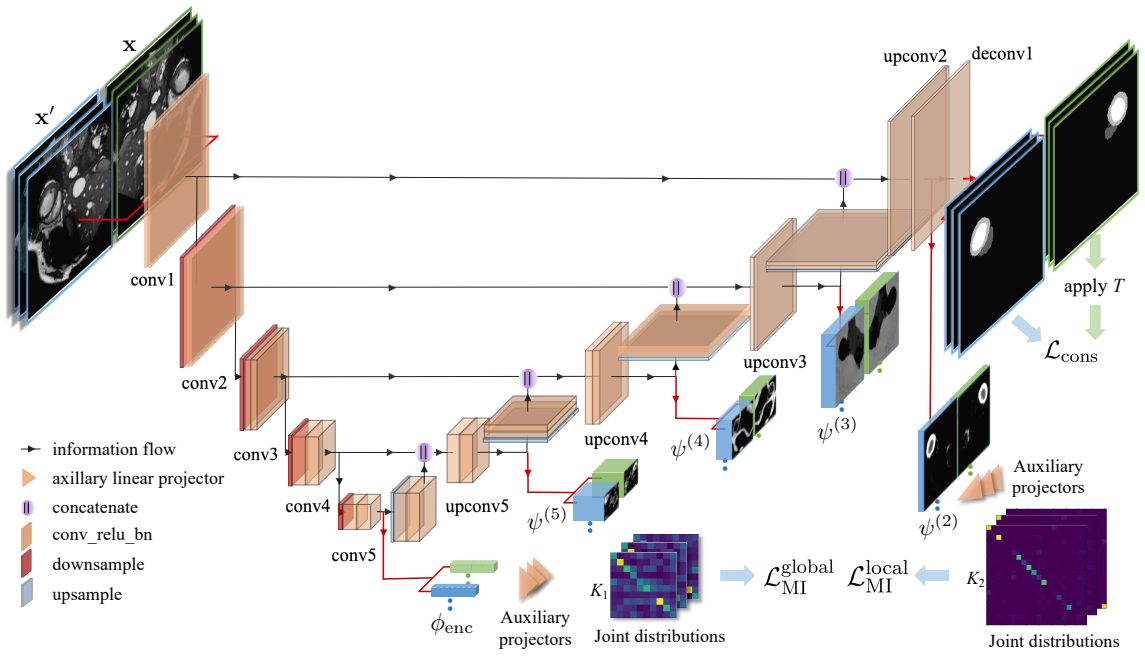

Figure 1: **Training pipeline of our semi-supervised segmentation method**. Given an unlabeled image $\mathbf{x}$ and its transformation $\mathbf{x}'$, we seek to maximize the mutual information of their intermediate feature representation with the help of auxiliary projectors. We maximize the global 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.

3. 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.

3.1 Semi-supervised segmentation model

We consider a semi-supervised segmentation task where we have a labeled dataset $\mathcal{D}_l$ of image-label pairs $(\mathbf{x}, \mathbf{y})$, with image $\mathbf{x} \in \mathbb{R}^{\Omega}$ and ground-truth labels $\mathbf{y} \in \{1, \dots, C\}^{\Omega}$, and a larger unlabeled dataset $\mathcal{D}_u$ consisting of images without their annotations. Here, $\Omega = \{1, \dots, W\} \times \{1, \dots, H\}$ 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 $\boldsymbol{\theta}$ to predict the segmentation label of each pixel of the input image.

Fig. 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 $\mathbf{x}$ by passing it through multiple convolutional blocks with down-sampling, and squeezes it into a compact embedding $\phi_{\text{enc}}(\mathbf{x})$. 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 $\mathbf{y} = \phi_{\text{dec}}(\phi_{\text{enc}}(\mathbf{x}))$. 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 $\boldsymbol{\theta}$ of the network are learned by optimizing the following loss function:

$$\mathcal{L}(\boldsymbol{\theta}; \mathcal{D}_l, \mathcal{D}_u) \;=\; \mathcal{L}_{\text{spv}}(\boldsymbol{\theta}; \mathcal{D}_l) \;+\; \lambda_1 \mathcal{L}_{\text{MI}}^{\text{global}}(\boldsymbol{\theta}; \mathcal{D}_u) \;+\; \lambda_2 \mathcal{L}_{\text{MI}}^{\text{local}}(\boldsymbol{\theta}; \mathcal{D}_u) \;+\; \lambda_3 \mathcal{L}_{\text{cons}}(\boldsymbol{\theta}; \mathcal{D}_u). \quad (3)$$

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:

$$\mathcal{L}_{\text{spv}}(\boldsymbol{\theta}; \mathcal{D}_l) \;=\; -\frac{1}{|\mathcal{D}_l|\,|\Omega|} \sum_{(\mathbf{x},\mathbf{y}) \in \mathcal{D}_l} \sum_{(i,j) \in \Omega} y_{ij} \log f_{ij}(\mathbf{x}; \boldsymbol{\theta}). \quad (4)$$

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 $\phi_{\text{enc}}(\mathbf{x})$ produced by the encoder as a global representation of an image $\mathbf{x}$, 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 (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.

3.2 Global mutual information loss

Let $\mathbf{x}$ 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 $\mathbf{x}$, the transformed image $\mathbf{x}' = T(\mathbf{x})$ should share similar contextual information as $\mathbf{x}$. 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:

$$\max_{\boldsymbol{\theta}_{\text{enc}}} \; I\big(\phi_{\text{enc}}(X);\, \phi_{\text{enc}}(X')\big) \tag{5}$$

where $\boldsymbol{\theta}_{\text{enc}}$ are the enconder's learnable parameters. However, optimizing directly Eq. (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 and Ermon, 2019).

To overcome this problem, we adapt the method proposed for unsupervised clustering and project the embeddings into categorical distributions $p(Z \,|\, \mathbf{x}) = g(\phi_{\text{enc}}(\mathbf{x}))) \in [0,1]^K$ with an auxiliary projector g consisting of a linear layer followed by a softmax activation. Using this approach, embeddings $\phi_{\text{enc}}(\mathbf{x})$ and $\phi_{\text{enc}}(\mathbf{x}')$ are converted to cluster probability distributions $p(Z \,|\, \mathbf{x})$ and $p(Z \,|\, \mathbf{x}')$ with a predefined cluster number K. This projection introduces a bottleneck effect on (5) since

$$I\big(g(\phi_{\text{enc}}(X));\, g(\phi_{\text{enc}}(X'))\big) \;\leq\; I\big(\phi_{\text{enc}}(X);\, \phi_{\text{enc}}(X')\big) \tag{6}$$

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 et al., 2000; Alemi et al., 2016; Ji et al., 2018). 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

$$p(Z, Z' \,|\, \mathbf{x}, \mathbf{x}') \;=\; g\big(\phi_{\text{enc}}(\mathbf{x})\big) \cdot g\big(\phi_{\text{enc}}(T(\mathbf{x}))\big)^{\top} \tag{7}$$

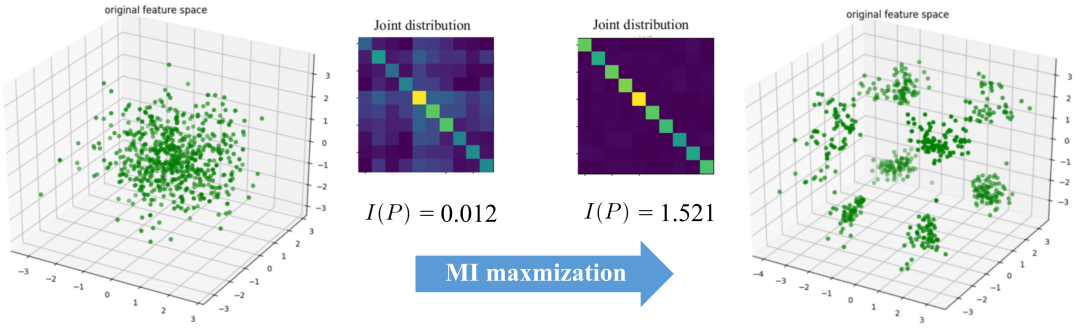

Figure 2: A toy example illustrating the effect of maximizing global MI. By increasing $I(P)$, randomly generated 3-D embedding points are effectively grouped into well-defined clusters.

yields a $K \times K$ probability matrix for each $\mathbf{x} \in \mathcal{D}_u$, $\mathbf{x}' = T(\mathbf{x})$, 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

$$P \approx \frac{1}{|\mathcal{D}_u|\,|\mathcal{T}|} \sum_{\mathbf{x}\in\mathcal{D}_u} \sum_{T\in\mathcal{T}} g\big(\phi_{\text{enc}}(\mathbf{x})\big) \cdot g\big(\phi_{\text{enc}}(T(\mathbf{x}))\big)^{\top}. \tag{8}$$

Using the definition of MI in (1), the proposed global MI loss can then computed from P as follows:

$$\mathcal{L}_{\text{MI}}^{\text{global}}(\boldsymbol{\theta}; \mathcal{D}_u) = -I(P) = -\sum_{c=1}^{K}\sum_{c'=1}^{K} P_{c,c'} \log \frac{P_{c,c'}}{\sum_{c=1}^{K} P_{c,c'} \cdot \sum_{c'=1}^{K} P_{c,c'}} \tag{9}$$

A high $I(P)$ means that the information of $T(\mathbf{x})$ can be retrieved given $\mathbf{x}$ (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. 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.

3.3 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 $\mathcal{L}_{\text{MI}}^{\text{local}}$ that preserves the local information of feature embeddings in the decoder.

Let $\psi^{(b)}(\mathbf{x}) = \phi_{\text{dec}}^{(b)}\big(\phi_{\text{enc}}(\mathbf{x})\big) \in \mathbb{R}^{C_b \times H_b \times W_b}$ be the feature map produced in the b-th decoder block for an unlabeled image $\mathbf{x}$. As described in Section 4.2, each block is composed of a convolution and an upsampling operation. This feature map has a reduced spatial resolution compared to $\mathbf{x}$ 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 $\psi^{(b)}(\mathbf{x})$. Denoting as $\big[\psi^{(b)}(\mathbf{x})\big]_{i,j} \in \mathbb{R}^C$ 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 $\Delta^{(b)} \subset \mathbb{Z}^2$:

$$\mathcal{N}_{i,j}^{(b)} \;=\; \big\{\big[\psi^{(b)}(\mathbf{x})\big]_{i+p,j+q} \,|\, (p,q) \in \Delta^{(b)}\big\}. \tag{10}$$

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 $\big[\psi^{(b)}(T(\mathbf{x}))\big]_{i,j}$ with $\big[T(\psi^{(b)}(\mathbf{x}))\big]_{i+p,j+q}$, where (p,q) is a displacement in $\Delta^{(b)}$. As before, we use a linear projection head h to convert the feature map $\psi^{(b)}(\mathbf{x})$ to a cluster assignment $h(\psi^{(b)}(\mathbf{x})) \in [0,1]^{K_b \times H_b \times W_b}$. 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., 2020; Ji et al., 2018), we then compute a separate joint distribution $P_{p,q}^{(b)}$ for each displacement $(p,q) \in \Delta^{(b)}$:

$$P_{p,q}^{(b)} \;\approx\; \frac{1}{|\mathcal{D}_u|\,|\mathcal{T}|\,|\Omega|} \sum_{\mathbf{x}\in\mathcal{D}_u} \sum_{T\in\mathcal{T}} \sum_{(i,j)\in\Omega} h\big([\psi^{(b)}(T(\mathbf{x}))]_{i,j}\big) \cdot h\big([T(\psi^{(b)}(\mathbf{x}))]_{i+p,j+q}\big)^\top \tag{11}$$

Note that the operation in (11) can be computed efficiently with standard convolution operations. Finally, we obtain the local MI loss by averaging the MI over all decoder blocks $b \in \{1,\dots,B\}$ and corresponding displacements:

$$\mathcal{L}_{\text{MI}}^{\text{local}}(\boldsymbol{\theta};\mathcal{D}_u) \;=\; -\frac{1}{B}\sum_{b=1}^{B} \frac{1}{|\Delta^{(b)}|} \sum_{(p,q)\in\Delta^{(b)}} I\big(P_{p,q}^{(b)}\big) \tag{12}$$

where $I\big(P_{p,q}^{(b)}\big)$ is computed as in (9).

3.4 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 $\mathcal{L}_{\text{cons}}$ 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:

$$\mathcal{L}_{\text{cons}}(\boldsymbol{\theta}; \mathcal{D}_u) \;=\; \frac{1}{|\mathcal{D}_u|\,|\mathcal{T}|\,|\Omega|} \sum_{\mathbf{x}\in\mathcal{D}_u} \sum_{T\in\mathcal{T}} \sum_{(i,j)\in\Omega} \left\| f_{ij}(T(\mathbf{x})) - T(f_{ij}(\mathbf{x})) \right\|_2^2 \tag{13}$$

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. Experimental setup

4.1 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 and Shen, 2016). These four datasets contain different image modalities (CT and MRI) and acquisition resolutions.

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 $1.0 \times 1.0\,$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.

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.

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.

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):

$$\mathrm{DSC}(S, G) = \frac{2|S \cap G|}{|S| + |G|} \tag{14}$$

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.

4.2 Implementation details

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. 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 (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.

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 $\mathcal{L}_{\text{cons}}$ 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.

5. Experimental Results

5.1 Comparison with the state-of-the-art

Table 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

Table 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)**

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 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. 3, 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.

5.2 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, the proposed method (Dark

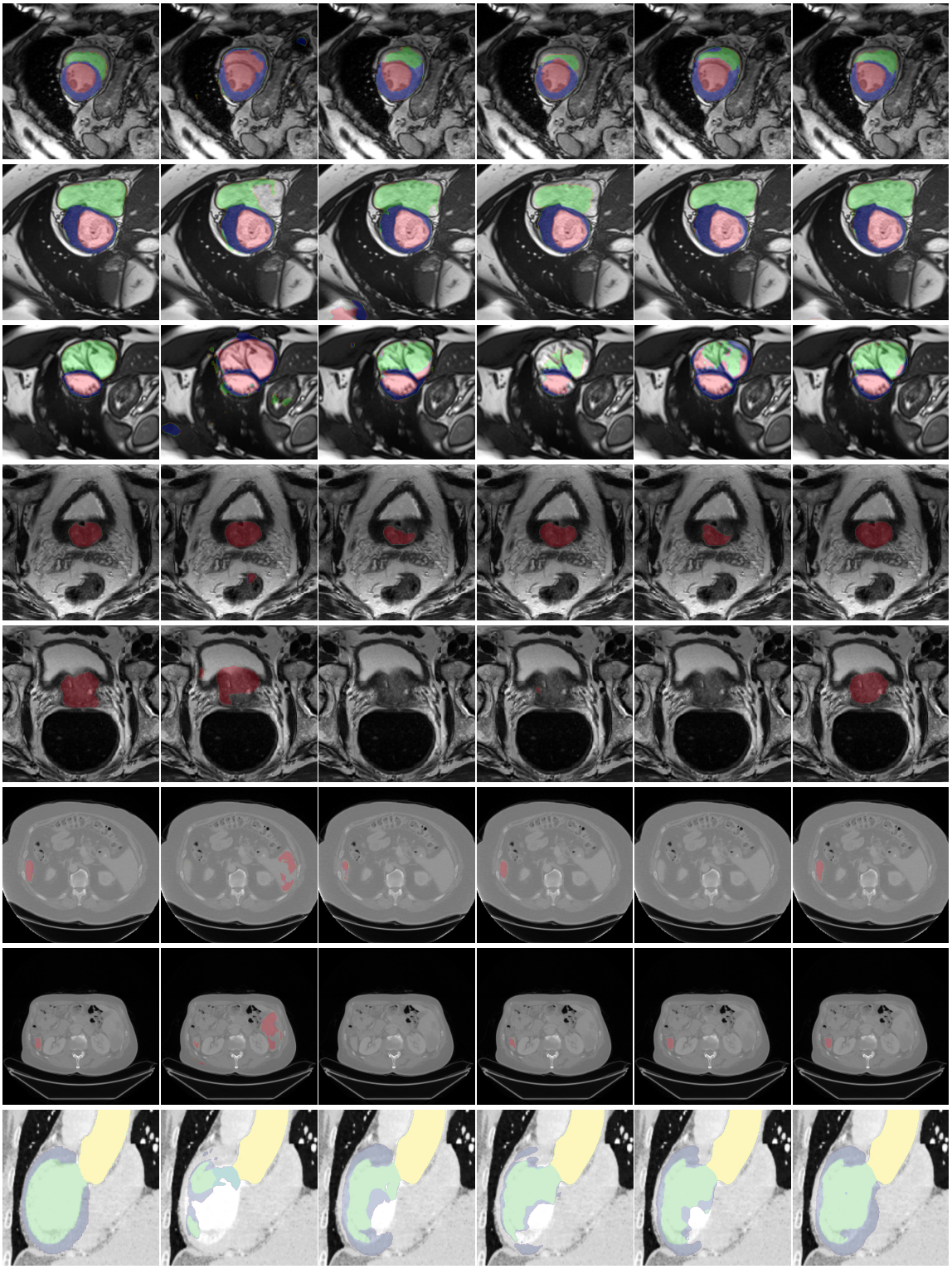

Figure 3: Visual comparison of tested methods on validation images. **Rows 1–3**: ACDC; **Rows 4–5**: Prostate; **Rows 6–7**: Spleen; **Row 8**: MMWHS.

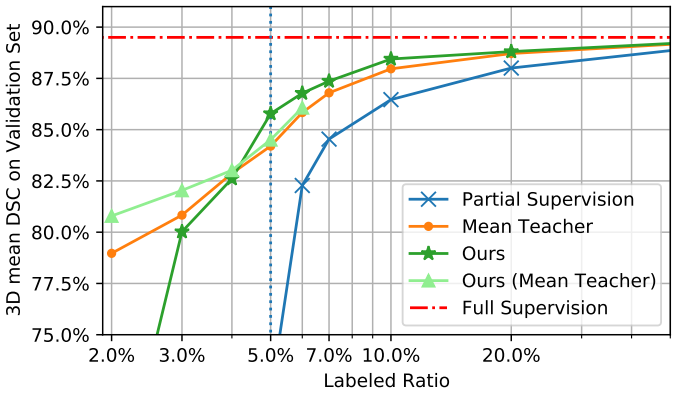

Figure 4: ACDC validation DSC versus various labeled data ratio for tested methods. It is clearly observed that our proposed method achieves higher performance compared with the state-of-the-art Mean Teacher in the regime of labeled ratio $\geq 5\%$. For an extreme case where only $2-3\%$ data are provided with annotations, our enhanced adaption outperforms Mean Teacher.

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.

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 (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.

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

Table 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**

Table 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%

Table 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.

5.4 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

Table 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%**

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 1×1 for both **Upconv5** and **Upconv4** as their resolution scales correspond to $1/8$ and $1/4$ of an input image.

We observe from Table 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.

5.5 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. 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 compu-

tational 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.

5.6 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. 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. 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. 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 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. 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.

5.7 Joint optimization of supervised and unsupervised losses

A significant difference between our method and (Ji et al., 2018) relates to how models are trained. The approach in (Ji et al., 2018) 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. (3), without the supervised loss term of Eq. (4). By doing so, the network tries to learn a meaningful feature representation without annotations. In the second stage,

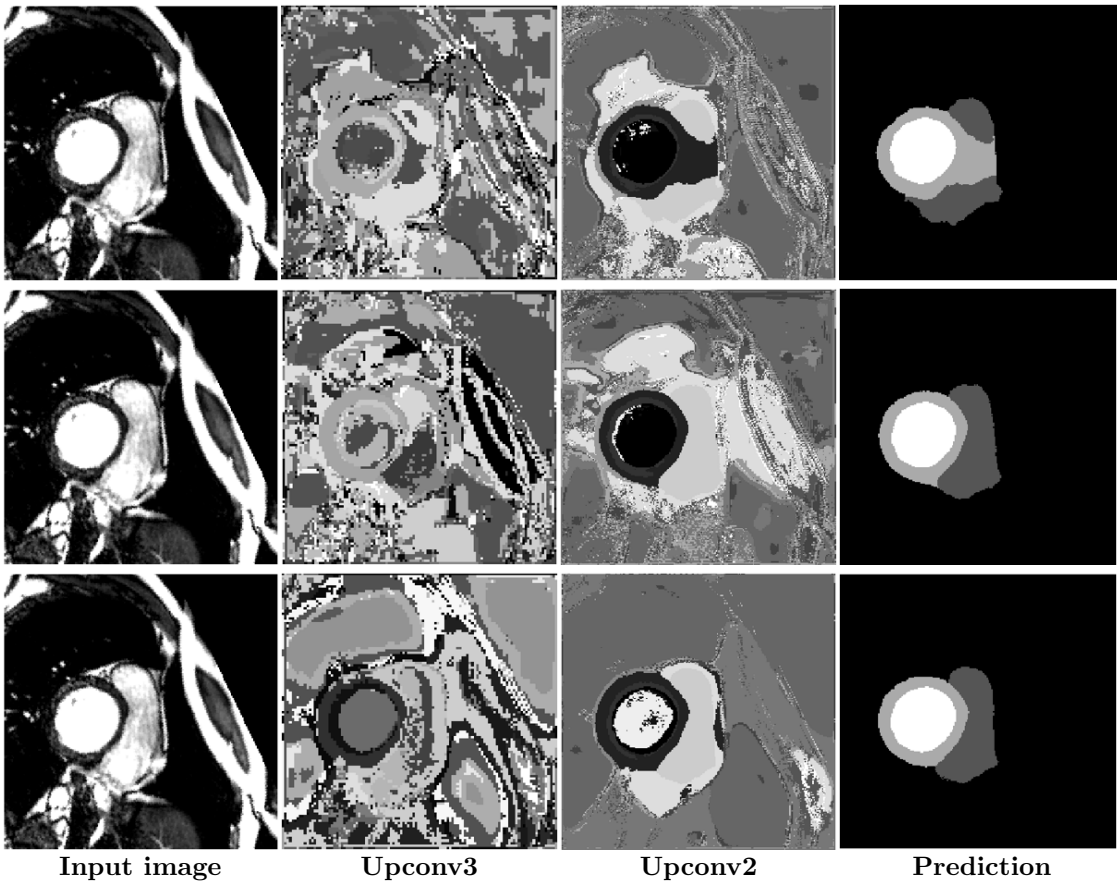

Figure 5: Visual comparison of maximum activations taken from network decoder positions. **Top row**: Partial Supervision. **Middle row**: Mean Teacher. **Bottom row**: Our method.

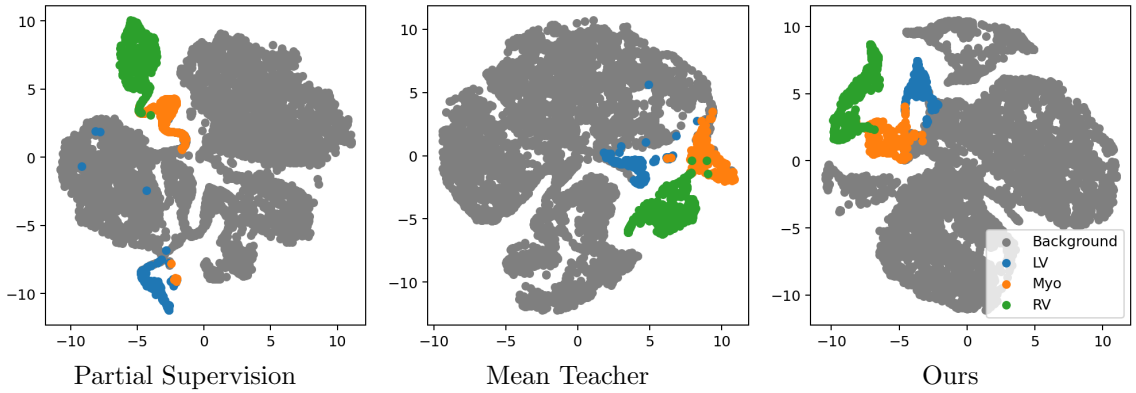

Figure 6: t-SNE plot on the ACDC validation set for different classes.

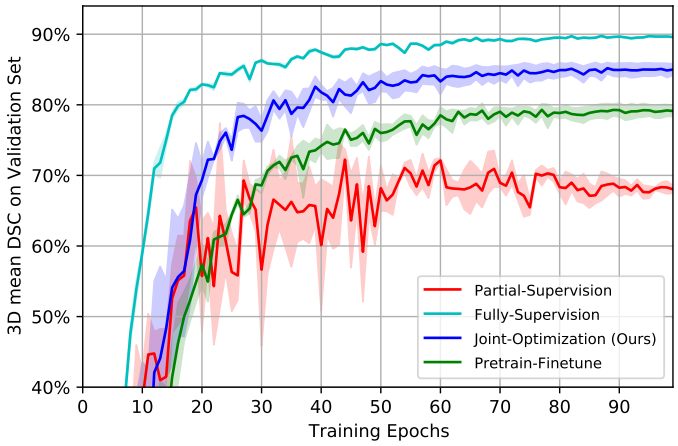

Figure 7: Validation mean DSC versus training epochs for the Pretrain-Finetune strategy and our Joint-Optimization method. Mean and standard deviation values are calculated from three independent runs.

we apply the supervised loss of Eq. (4) 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. 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.

6. 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 et al., 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., 2017). 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).

Acknowledgments

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), and thank NVIDIA corporation for supporting this work through their GPU grant program.

Ethical Standards

The work follows appropriate ethical standards in conducting research and writing the manuscript, following all applicable laws and regulations regarding treatment of animals or human subjects.

Conflicts of Interest

We declare we don't have conflicts of interest.

References

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.

Wenjia Bai, Ozan Oktay, Matthew Sinclair, Hideaki Suzuki, Martin Rajchl, Giacomo Tarroni, Ben Glocker, Andrew King, Paul M Matthews, and Daniel Rueckert. Semi-supervised learning for network-based cardiac MR image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 253–260. Springer, 2017.

Christoph Baur, Shadi Albarqouni, and Nassir Navab. Semi-supervised deep learning for fully convolutional networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 311–319. Springer, 2017.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. MINE: mutual information neural estimation. *arXiv preprint arXiv:1801.04062*, 2018.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved? *IEEE Transactions on Medical Imaging*, 2018.

Gerda Bortsova, Florian Dubost, Laurens Hogeweg, Ioannis Katramados, and Marleen de Bruijne. Semi-supervised medical image segmentation via learning consistency under transformations. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 810–818. Springer, 2019.

Avishek Joey Bose, Huan Ling, and Yanshuai Cao. Adversarial contrastive estimation. *arXiv preprint arXiv:1805.03642*, 2018.

John S Bridle, Anthony JR Heading, and David JC MacKay. Unsupervised classifiers, mutual information and'phantom targets. In *Advances in neural information processing systems*, pages 1096–1101, 1992.

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 132–149, 2018.

Krishna Chaitanya, Neerav Karani, Christian F Baumgartner, Anton Becker, Olivio Donati, and Ender Konukoglu. Semi-supervised and task-driven data augmentation. In *International Conference on Information Processing in Medical Imaging*, pages 29–41. Springer, 2019.

Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu. Contrastive learning of global and local features for medical image segmentation with limited annotations. *arXiv preprint arXiv:2006.10511*, 2020.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

Wenhui Cui, Yanlin Liu, Yuxing Li, Menghao Guo, Yiming Li, Xiuli Li, Tianle Wang, Xiangzhu Zeng, and Chuyang Ye. Semi-supervised brain lesion segmentation with an adapted mean teacher model. In *International Conference on Information Processing in Medical Imaging*, pages 554–565. Springer, 2019.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

Weihua Hu, Takeru Miyato, Seiya Tokui, Eiichi Matsumoto, and Masashi Sugiyama. Learning discrete representations via information maximizing self-augmented training. *ICML*, 2017.

Xu Ji, João F. Henriques, and Andrea Vedaldi. Invariant information distillation for unsupervised image segmentation and clustering. *CoRR*, abs/1807.06653, 2018. URL `http://arxiv.org/abs/1807.06653`.

Dahun Kim, Donghyeon Cho, Donggeun Yoo, and In So Kweon. Learning image representations by completing damaged jigsaw puzzles. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 793–802. IEEE, 2018.

Andreas Krause, Pietro Perona, and Ryan G. Gomes. Discriminative clustering by regularized information maximization. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 775–783. Curran Associates, Inc., 2010. URL `http://papers.nips.cc/paper/4154-discriminative-clustering-by-regularized-information-maximization.pdf`.

Xiaomeng Li, Lequan Yu, Hao Chen, Chi-Wing Fu, and Pheng-Ann Heng. Semi-supervised skin lesion segmentation via transformation consistent self-ensembling model. *arXiv preprint arXiv:1808.03887*, 2018.

Ruizhi Liao, Daniel Moyer, Polina Golland, and William M Wells. Demi: Discriminative estimator of mutual information. *arXiv preprint arXiv:2010.01766*, 2020.

Geert Litjens, Robert Toth, Wendy van de Ven, Caroline Hoeks, Sjoerd Kerkstra, Bram van Ginneken, Graham Vincent, Gwenael Guillard, Neil Birbeck, Jindang Zhang, et al. Evaluation of prostate segmentation algorithms for MRI: the PROMISE12 challenge. *Medical image analysis*, 18(2):359–373, 2014.

David McAllester and Karl Stratos. Formal limitations on the measurement of mutual information. In *International Conference on Artificial Intelligence and Statistics*, pages 875–884. PMLR, 2020.

Shaobo Min and Xuejin Chen. A robust deep attention network to noisy labels in semi-supervised biomedical segmentation. *arXiv preprint arXiv:1807.11719*, 2018.

T. Miyato, S. Maeda, M. Koyama, and S. Ishii. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1979–1993, Aug 2019. ISSN 0162-8828. doi: 10.1109/TPAMI.2018.2858821.

Sudipto Mukherjee, Himanshu Asnani, and Sreeram Kannan. Ccmi: Classifier based conditional mutual information estimation. In *Uncertainty in Artificial Intelligence*, pages 1083–1093. PMLR, 2020.

Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*, pages 69–84. Springer, 2016.

Mehdi Noroozi, Ananth Vinjimoor, Paolo Favaro, and Hamed Pirsiavash. Boosting self-supervised learning via knowledge transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9359–9367, 2018.

Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In *Advances in neural information processing systems*, pages 271–279, 2016.

Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 3239–3250. Curran Associates, Inc., 2018.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Jizong Peng, Christian Desrosiers, and Marco Pedersoli. Information based deep clustering: An experimental study. *arXiv preprint arXiv:1910.01665*, 2019a.

Jizong Peng, Guillermo Estradab, Marco Pedersoli, and Christian Desrosiers. Deep co-training for semi-supervised image segmentation. *arXiv preprint arXiv:1903.11233*, 2019b.

Jizong Peng, Marco Pedersoli, and Christian Desrosiers. Mutual information deep regularization for semi-supervised segmentation. *Proceedings of Machine Learning Research*, 1: 13, 2020.

Christian S Perone and Julien Cohen-Adad. Deep semi-supervised segmentation with weight-averaged consistency targets. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 12–19. Springer, 2018.

Christian S Perone, Pedro Ballester, Rodrigo C Barros, and Julien Cohen-Adad. Unsupervised domain adaptation for medical imaging segmentation with self-ensembling. *NeuroImage*, 194:1–11, 2019.

Ben Poole, Sherjil Ozair, Aaron van den Oord, Alexander A Alemi, and George Tucker. On variational bounds of mutual information. *arXiv preprint arXiv:1905.06922*, 2019.

Ilija Radosavovic, Piotr Dollár, Ross Girshick, Georgia Gkioxari, and Kaiming He. Data distillation: Towards omni-supervised learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4119–4128, 2018.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

Amber L Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram van Ginneken, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063*, 2019.

Jiaming Song and Stefano Ermon. Understanding the limitations of variational mutual information estimators. *arXiv preprint arXiv:1910.06222*, 2019.

Nasim Souly, Concetto Spampinato, and Mubarak Shah. Semi supervised semantic segmentation using generative adversarial network. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 5689–5697. IEEE, 2017.

Taiji Suzuki, Masashi Sugiyama, Jun Sese, and Takafumi Kanamori. Approximating mutual information by maximum likelihood density ratio estimation. In *New challenges for feature selection in data mining and knowledge discovery*, pages 5–20, 2008.

Aiham Taleb, Christoph Lippert, Tassilo Klein, and Moin Nabi. Multimodal self-supervised learning for medical image analysis. *arXiv preprint arXiv:1912.05396*, 2019.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Martin Vejmelka and Kateřina Hlaváčková-Schindler. Mutual information estimation in higher dimensions: A speed-up of a k-nearest neighbor based estimator. In *International conference on adaptive and natural computing algorithms*, pages 790–797. Springer, 2007.

Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2517–2526, 2019.

Xiao Wang, Daisuke Kihara, Jiebo Luo, and Guo-Jun Qi. Enaet: Self-trained ensemble autoencoding transformations for semi-supervised learning. *arXiv preprint arXiv:1911.09265*, 2019.

Lequan Yu, Shujun Wang, Xiaomeng Li, Chi-Wing Fu, and Pheng-Ann Heng. Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 605–613. Springer, 2019.

Liheng Zhang, Guo-Jun Qi, Liqiang Wang, and Jiebo Luo. Aet vs. aed: Unsupervised representation learning by auto-encoding transformations rather than data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2547–2555, 2019.

Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016.

Yizhe Zhang, Lin Yang, Jianxu Chen, Maridel Fredericksen, David P Hughes, and Danny Z Chen. Deep adversarial networks for biomedical image segmentation utilizing unannotated images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 408–416. Springer, 2017.

Amy Zhao, Guha Balakrishnan, Fredo Durand, John V Guttag, and Adrian V Dalca. Data augmentation using learned transformations for one-shot medical image segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8543–8553, 2019a.

Junjie Zhao, Donghuan Lu, Kai Ma, Yu Zhang, and Yefeng Zheng. Deep image clustering with category-style representation. In *European Conference on Computer Vision*, pages 54–70. Springer, 2020.

Shuai Zhao, Yang Wang, Zheng Yang, and Deng Cai. Region mutual information loss for semantic segmentation. In *Advances in Neural Information Processing Systems*, pages 11115–11125, 2019b.

Yuyin Zhou, Yan Wang, Peng Tang, Song Bai, Wei Shen, Elliot Fishman, and Alan Yuille. Semi-supervised 3D abdominal multi-organ segmentation via deep multi-planar co-training. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 121–140. IEEE, 2019.

Xiahai Zhuang and Juan Shen. Multi-scale patch and multi-modality atlases for whole heart segmentation of MRI. *Medical image analysis*, 31:77–87, 2016.