Medical Image Analysis 87 (2023) 102792

Contents lists available at ScienceDirect

Medical Image Analysis

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

Local contrastive loss with pseudo-label based self-training for semi-supervised medical image segmentation

Krishna Chaitanya *, Ertunc Erdil, Neerav Karani, Ender Konukoglu

Computer Vision Laboratory, ETH Zurich, Sternwartstrasse 7, Zurich 8092, Switzerland

ARTICLE INFO

Keywords:
Contrastive learning
Self-supervised learning
Semi-supervised learning
Machine learning
Deep learning
Medical image segmentation

ABSTRACT

Supervised deep learning-based methods yield accurate results for medical image segmentation. However, they require large labeled datasets for this, and obtaining them is a laborious task that requires clinical expertise. Semi/self-supervised learning-based approaches address this limitation by exploiting unlabeled data along with limited annotated data. Recent self-supervised learning methods use contrastive loss to learn good global level representations from unlabeled images and achieve high performance in classification tasks on popular natural image datasets like ImageNet. In pixel-level prediction tasks such as segmentation, it is crucial to also learn good local level representations along with global representations to achieve better accuracy. However, the impact of the existing local contrastive loss-based methods remains limited for learning good local representations because similar and dissimilar local regions are defined based on random augmentations and spatial proximity; not based on the semantic label of local regions due to lack of large-scale expert annotations in the semi/self-supervised setting. In this paper, we propose a local contrastive loss to learn good pixel level features useful for segmentation by exploiting semantic label information obtained from pseudo-labels of unlabeled images alongside limited annotated images with ground truth (GT) labels. In particular, we define the proposed contrastive loss to encourage similar representations for the pixels that have the same pseudo-label/GT label while being dissimilar to the representation of pixels with different pseudo-label/GT label in the dataset. We perform pseudo-label based self-training and train the network by jointly optimizing the proposed contrastive loss on both labeled and unlabeled sets and segmentation loss on only the limited labeled set. We evaluated the proposed approach on three public medical datasets of cardiac and prostate anatomies, and obtain high segmentation performance with a limited labeled set of one or two 3D volumes. Extensive comparisons with the state-of-the-art semi-supervised and data augmentation methods and concurrent contrastive learning methods demonstrate the substantial improvement achieved by the proposed method. The code is made publicly available at https://github.com/krishnabits001/pseudo_label_contrastive_training.

1. Introduction

Medical image segmentation with high accuracy is very desirable for many downstream clinical applications. Currently, supervised deep learning methods yield state-of-the-art segmentation performance (Ronneberger et al., 2015; Milletari et al., 2016; Kamnitsas et al., 2016, 2017), however, this approach requires having access to large labeled datasets. Obtaining such large labeled datasets for segmentation is time-consuming and expensive in medical imaging since the annotations need to be done by clinical experts. To attenuate the need for large labeled datasets, semi-supervised (Lee et al., 2013; Rasmus et al., 2015) and self-supervised learning methods (Doersch et al., 2015; Chen et al., 2020a) leverage unlabeled data along with limited labeled examples. Recent methods have yielded promising results on medical datasets and reduced the requirement for a large number of the labeled examples (Bai et al., 2017; Chaitanya et al., 2020).

1.1. Motivation

Prominent works in semi-supervised learning methods focus on extracting useful information from unlabeled examples along with using limited labeled examples. Some popular semi-supervised learning strategies include using pseudo-labels for self-training (Lee et al., 2013; Bai et al., 2017), entropy minimization (Grandvalet and Bengio, 2005), consistency regularization (Sajjadi et al., 2016; Laine and Aila,

* Corresponding author.
E-mail address: krishna.chaitanya@vision.ee.ethz.ch (K. Chaitanya).

https://doi.org/10.1016/j.media.2023.102792
Received 24 November 2021; Received in revised form 25 November 2022; Accepted 2 March 2023
Available online 11 March 2023

2016) and data augmentation (Chaitanya et al., 2019; Zhao et al., 2019). Similarly, self-supervised learning methods such as pretext-task-based (Doersch et al., 2015) and contrastive learning-based (Chen et al., 2020a) methods aim to learn a good network initialization via pre-training with only unlabeled images and later fine-tune this initialization with limited annotations to get high performance.

Contrastive learning-based methods yielded highly accurate models on many natural and medical image datasets for classification and segmentation tasks. The objective of contrastive learning (Hadsell et al., 2006) is that the latent representations of similar images should be alike, and simultaneously they should be different from the representations of the dissimilar images. Therefore, it is crucial to correctly define similar and dissimilar images during optimization to obtain better representations. Earlier works (Chen et al., 2020a) define similar images as two differently transformed views of an image while defining dissimilar images as images with different contents. Some recent methods (Chaitanya et al., 2020) define similarity cues by using domain knowledge for medical images that go beyond random transformations by defining an additional set of similar images as slices from corresponding anatomical areas across subjects. These methods learn global-level image representations useful for downstream tasks such as classification and segmentation and achieve high performance when fine-tuned with limited labels.

Unsupervised learning of good local features can be as crucial as global features for pixel-level prediction tasks such as segmentation. However, the performance of local contrastive learning-based methods is rather limited compared to global ones due to the difficulty of defining similar/dissimilar local regions without semantic labels. Early work by Chaitanya et al. (2020) enforce patch-level local features across different augmented views of an image to be similar to each other while being dissimilar to representations of other local regions within the image. More recent local contrastive learning methods extend this work by using surrogate semantic labels (Hénaff et al., 2021; Van Gansbeke et al., 2021) where the labels are estimated on unlabeled images using unsupervised techniques such as saliency maps estimation approaches (Nguyen et al., 2019; Pinheiro et al., 2015; Arbelaez et al., 2010), super-pixels (Achanta et al., 2012), and image computable masks (Felzenszwalb and Huttenlocher, 2004; Arbeláez et al., 2014). These methods encourage local regions having the same semantic label to have similar representations while being dissimilar to regions with different labels. There are two possible shortcomings of using unsupervised methods to obtain surrogate segmentation masks of unlabeled images: (1) surrogate class labels obtained by an unsupervised method most likely do not match with the desired target semantic classes, (2) similarity/dissimilarity of local regions across different images cannot be enforced since these methods do not necessarily assign the same label id for the same anatomy across different images. This is illustrated in Fig. 1. These problems may limit the quality of the learned representations. Also, these approaches consist of two stages training: pre-training with surrogate labels and fine-tuning with limited target labels for any downstream tasks like classification or segmentation. This motivates us to formulate an end-to-end joint training framework suited for segmentation task where we devise to use pseudo-labels of unlabeled data instead of surrogate labels from unsupervised methods to overcome these problems and achieve high segmentation accuracy.

1.2. Contributions

In this paper, we address the above limitations by proposing an end-to-end joint training semi-supervised approach more suitable for segmentation task by defining a local pixel-level contrastive loss on pseudo-labels of unlabeled set and limited labeled set with ground truth labels. The proposed contrastive loss encourages similar representations for the pixels that belong to the same target semantic label while simultaneously enforcing them to be dissimilar to the representations of the pixels from different classes. In this article, when we state "limited labeled set", we refer to the ground-truth labels of the labeled set.

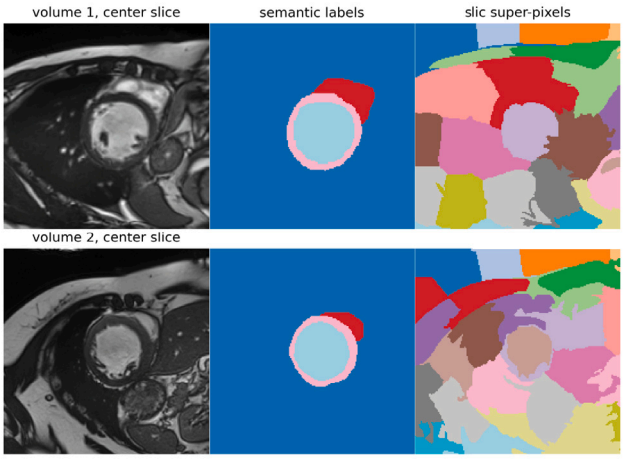

Fig. 1. A visual example that demonstrates the shortcoming of existing local contrastive learning methods. Unsupervised segmentation methods such as a super-pixel based method Slic (Achanta et al., 2012) (1) do not produce semantic labels similar to ground truth, (2) do not assign the same label id for the same anatomy across two different images (see different colors assigned by Slic for the similar regions) which hinders to enforce similarity/dissimilarity of pixel representations of same anatomies across images (Hénaff et al., 2021).

Our contributions in this article are stated below:

(a) The contrastive loss objective enforces intra-class similarity and inter-class separability for the representations of target semantic classes from both ground-truths of labeled images and pseudo-labels of unlabeled images across the dataset desirable for the segmentation task, which cannot be enforced or modeled in the prior semi-supervised methods (Bai et al., 2017; Verma et al., 2019; Luo et al., 2021; Chen et al., 2021) that are based on cross-entropy loss with pseudo-labels. This is because cross-entropy loss cannot match same class representations across images due to its definition. We observe the gains from using such contrastive loss formulation over cross-entropy formulation as shown in Table 1.

(b) Unlike concurrent semi-supervised contrastive learning works (Zhao et al., 2020; Alonso et al., 2021), we do not use the pseudo-labels of unlabeled images in the cross-entropy loss directly as using them directly can lead to reinforcing erroneous predictions in the training (Chapelle et al., 2009). Here, we instead have two separate network branches to compute segmentation and contrastive loss, where the segmentation branch only considers labeled set and the contrastive branch considers both labeled set and unlabeled set with their pseudo-labels as shown in Fig. 2. In addition, we do not perform two-step training. Instead, propose to use a single step joint training. We observe the performance benefits of this design as presented in Table 1.

(c) To further mitigate erroneous prediction's effect on network training, we additionally considered using better pre-trained initialization (Chaitanya et al., 2020) and using consistency loss (Laine and Aila, 2016; Bortsova et al., 2019) to use only confidently predicted pseudo-labels in contrastive loss formulation beyond separate segmentation specific layers. We observed that good pre-trained initialization improves the performance of both the baselines and our proposed approach as shown in Table 2.

(d) We evaluated the benefits of the proposed method on three public MRI medical datasets and compared the performance with state-of-the-art self-supervised learning, semi-supervised learning, and concurrent contrastive learning methods. The results demonstrate that the proposed method yields higher gains than the best performing methods.

(e) We perform a detailed ablation study to understand how the performance is affected by each component or loss term defined in the learning framework.

2. Related work

We broadly classify the relevant literature into below categories with respect to the proposed method:

1. Self-supervised learning (SSL): Many recent works using SSL with appropriate unsupervised loss have shown to learn useful representations from unlabeled data. Models obtained from such optimization serve as good initialization for various downstream tasks and yield performance gains over random initialization. SSL works relevant to the presented work can be categorized into below two categories:

(i) Pre-text task-based methods: Here, a pre-text task is devised whose labels can be acquired with ease and freely from the unlabeled data itself to learn the initialization. Some examples of such tasks include: predicting the rotation applied on the input image (Gidaris et al., 2018), in-painting the missing pixel values (Pathak et al., 2016), context restoration (Chen et al., 2019), and many more (Doersch et al., 2015; Noroozi and Favaro, 2016; Zhang et al., 2016; Dosovitskiy et al., 2014).

(ii) Contrastive learning methods: These methods (Wu et al., 2018; Oord et al., 2018; Hjelm et al., 2019; Chen et al., 2020a; He et al., 2019; Hénaff et al., 2019; Misra and van der Maaten, 2019) use a contrastive loss (Hadsell et al., 2006; Gutmann and Hyvärinen, 2010) to enforce the representations from positive pairs to be similar and simultaneously be dissimilar to other representations. In general, positive pairs are defined by applying two different transformations (random augmentations) on an image. While some recent approaches use the domain-specific knowledge to capture more complex similarity cues over random augmentations by defining additional similar and dissimilar pairs, which have shown improved results for videos (Tschannen et al., 2020) and medical images (Chaitanya et al., 2020) over just using random transformations on an image. Most of these methods are defined to learn good global level representations that are useful for classification tasks. For dense prediction tasks such as segmentation, there are some relevant works which we list into below three categories:

(a) Supervised local contrastive learning: In these works (Zhao et al., 2020; Wang et al., 2021; Lee et al., 2021) authors have access to many labeled examples which they use to define an additional contrastive loss over standard segmentation loss to learn discriminative local pixel level features for the similar semantic classes over the whole dataset. One work (Hu et al., 2021) explores this supervised approach on medical image segmentation with limited annotations and get some improvements. In this work, we focus on the scenario where only limited labels are available.

(b) Unsupervised local contrastive learning: Here, the authors do not have access to ground-truth masks. Some methods (Chaitanya et al., 2020; Xie et al., 2020b; Wang et al., 2020; Jaus et al., 2021) pre-train by matching patch/pixel-level representations across either two augmented views of an image or a positive pair of images defined using domain cues. Other methods (Van Gansbeke et al., 2021; Hénaff et al., 2021) obtain surrogate masks of fore-ground objects using unsupervised techniques such as saliency maps (Nguyen et al., 2019; Pinheiro et al., 2015; Arbelaez et al., 2010), super-pixels (Achanta et al., 2012), image computable masks (Felzenszwalb and Huttenlocher, 2004; Arbeláez et al., 2014). Later, the local level features of a chosen fore-ground object are optimized to be similar while other class objects local features are optimized to be dissimilar. Here, the similar pixels or regions are chosen for the same object across two different transformations applied on an image. Often the fore-ground object classes estimated via the surrogate masks may not resemble with any of the multiple target classes to be segmented in medical images as shown in Fig. 1. Also, these methods like Slic (Achanta et al., 2012) do not generate consistent label ids for similar structures across different subjects as shown in Fig. 1. Therefore, in training, we cannot capture the complex similarity cues available for the same class across different subjects due to lack of true labels.

(c) Semi-supervised local contrastive learning: Recently, some concurrent works (Alonso et al., 2021; Zhao et al., 2020; Peng et al., 2021a; You et al., 2021a; Xiang et al., 2021; Zhou et al., 2021; You et al., 2021b) have devised semi-supervised learning frameworks that use unlabeled images with some variant of contrastive loss setup. The works relevant to presented work are (Alonso et al., 2021; Zhao et al., 2020). In Alonso et al. (2021), they use the pseudo-labels of unlabeled images in both cross-entropy loss and contrastive loss while (Zhao et al., 2020) use a two step iterative training process: pre-training with contrastive loss and later fine-tuning with both labeled annotations and pseudo-labels of unlabeled set for the downstream task. We only use the pseudo-labels of unlabeled images for contrastive loss computation and avoid using them in the segmentation loss (e.g., cross entropy or Dice loss) computation. We train the network in one step with a joint training setup using both contrastive and cross-entropy losses simultaneously. Also, in Alonso et al. (2021) they use a student–teacher network, employ entropy minimization on unlabeled set predictions, and use a separate memory bank to store and update the feature representations of the each class used in the training. We propose a much simpler setup with only one network and do not use any memory bank or entropy minimization. In limited annotation settings, where pseudo-labels are not very accurate, we observe empirically in our experiments that using them directly in cross-entropy loss hinders the network from obtaining higher performance gains. We observe the same for the two step training method. These results are presented in Table 1. Similar to Alonso et al. (2021), some other recent/concurrent works (Zhou et al., 2021; You et al., 2021b) also apply such contrastive loss with the teacher and student networks. In Zhou et al. (2021), they apply a pixel-wise contrastive loss on highly confident predicted regions from the same class, and additionally leverage consistency loss between predictions of the teacher and student networks. In You et al. (2021b), they apply the contrastive loss alongside segmentation and other consistency losses. Here, the contrastive loss is designed to learn object shape information with the help of boundary-aware representations, defined on the predicted signed distance maps of teacher and student networks.

2. Semi-supervised learning: There has been enormous amount of works proposed in semi-supervised learning that leverage unlabeled and labeled images together for training. Below we only describe three sets of works relevant to the proposed work.

(a) Self-training: In this method (Lee et al., 2013), an initial set of predictions are estimated for unlabeled images from a network trained with limited annotated labeled set. Later, the network is trained in an iterative fashion using both labeled images annotations and unlabeled images predictions (pseudo-labels) as proxy ground truths. The pseudo-labels estimates are updated after every few epochs of training and we expect their quality to improve through the training. This have shown improvements for medical image segmentation (Bai et al., 2017; Fan et al., 2020). But it has been found (Chapelle et al., 2009) that if the initial pseudo-label estimates are erroneous, then using them directly in the segmentation loss function can lead to possible degradation of performance. To avoid this, some methods (Nair et al., 2020; Yu et al., 2019a; Graham et al., 2019; Jungo and Reyes, 2019; Cao et al., 2020; Mehrtash et al., 2020; Camarasa et al., 2020) integrate uncertainty or confidence estimates (Blundell et al., 2015; Gal and Ghahramani, 2016; Kendall and Gal, 2017) of pseudo-labels into self-training to control the quality of pseudo-labels used for training and thereby reduce the negative effects of poor quality of pseudo-labels.

Another method that uses this idea is Noisy Student (Xie et al., 2020a), where two separate models called teacher-student networks are used. The teacher model is used to estimate pseudo-labels while the student model is trained with ground-truths of labeled set and pseudo-labels of the unlabeled set with consistency loss applied over random augmentations, dropout, and stochastic depth. Here, the teacher model is replaced with the latest student model after a pre-defined number of iterations to estimate refined pseudo-labels.

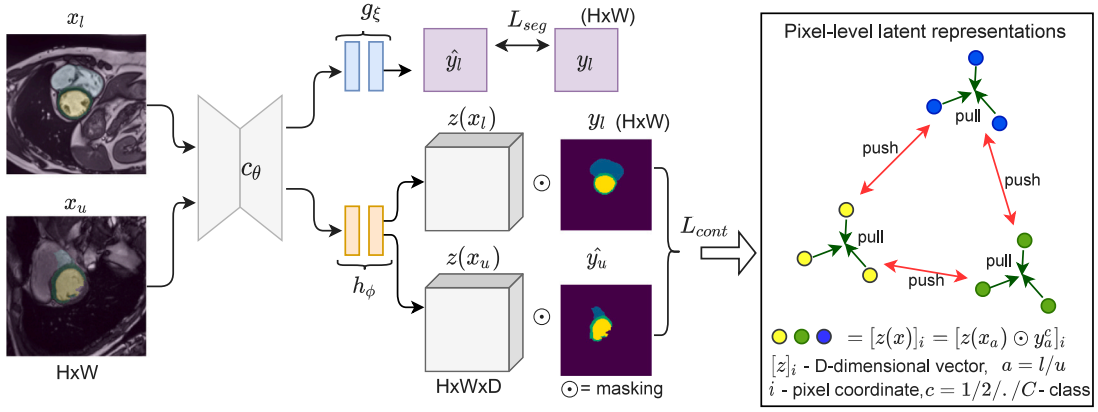

Fig. 2. The figure presents the proposed semi-supervised method to get high segmentation performance by applying a pixel-level contrastive loss on pseudo-labels of unlabeled data $(X_U, \hat{Y_U})$ and limited labeled data (X_L, Y_L) along with segmentation loss applied only on labeled set. The pixel-level representations of same class are optimized to be similar and simultaneously optimized to be dissimilar to other class representations in the whole dataset (shown on the right side). Here, c_θ is the backbone encoder–decoder network with two separate smaller network branches after it, where: (i) g_ξ predicts segmentation mask $(\hat{y}_l)$ used to compute the segmentation loss, and (ii) h_ϕ predicts the feature map of labeled image $(z(x_l)$ and unlabeled image $(z(x_u))$ that are used to compute the contrastive loss using the ground-truth mask (y_l) of labeled image x_l and pseudo-label mask $(\hat{y_u})$ of unlabeled image x_u.

(b) Consistency regularization: Such methods (Sajjadi et al., 2016; Laine and Aila, 2016; Tarvainen and Valpola, 2017) rely on the presumption that different perturbations/data augmented versions of the same image should yield same output label. It is expected that the network retains consistency in the output irrespective of changes made to the same image. The output label distribution loss is minimized with either mean square error or KL divergence loss between the perturbed/augmented samples generated. Some popular methods that apply this include pi-model (Sajjadi et al., 2016), temporal ensembling (Laine and Aila, 2016), mean-teacher (Tarvainen and Valpola, 2017), virtual adversarial training (Miyato et al., 2018). Following works (Bortsova et al., 2019; Cui et al., 2019; Yu et al., 2019a; Li et al., 2020a; Zhou et al., 2020; Fotedar et al., 2020; Peng et al., 2020; Fang and Li, 2020) apply some variants and combinations of the above ideas for medical image segmentation. Other recent works that apply consistency regularization on predictions of unlabeled images include (Verma et al., 2019; Luo et al., 2021; Chen et al., 2021). The two later works additionally leverage the signed distance map (SDM) that acts as a global shape constraint and regularizes the training. Similarly, in a concurrent work (Seibold et al., 2021), they apply consistency regularization with data augmentation and have an additional nearest neighbor computation on the features of unlabeled examples to compute their class pseudo-labels where the labeled examples act as reference nearest neighbors. This is done to mitigate relying directly on confidently predicted incorrect pseudo-labels from the network.

(c) Other methods include adversarial training with GANs (Zhang et al., 2017b; Nie et al., 2018; Zhang et al., 2018; Zheng et al., 2019; Han et al., 2020; Li et al., 2020b; Valvano et al., 2021), entropy minimization (Grandvalet and Bengio, 2005), and hybrid methods (Berthelot et al., 2019b,a; Sohn et al., 2020) which use a combination of above methods along with additional regularization terms. Some methods (Vu et al., 2019; You et al., 2020) leverage the domain adaptation setup. Lastly, in You et al. (2022) the authors use a transformer in an adversarial framework and rely on its capabilities to learn better high and low level features of semantic structures relevant for segmentation.

3. Data Augmentation: Prior works have shown that using affine data augmentations (Cireşan et al., 2011) such as crop, scale, rotation, flip and other augmentations such as random elastic deformations (Ronneberger et al., 2015), contrast/brightness based intensity augmentations (Hong et al., 2017; Perez et al., 2018) improves the baseline for medical imaging tasks. Also, MixUp augmentations (Zhang et al., 2017a; Eaton-Rosen et al., 2018) has shown to yield benefits on medical datasets. All the above stated methods only work with labeled examples and do not leverage unlabeled images. Other works have used generative adversarial networks (GANs) to generate additional synthetic image-label pairs (Goodfellow et al., 2014; Costa et al., 2018; Shin et al., 2018; Bowles et al., 2018; Cai et al., 2019; Yu et al., 2019b; Qin et al., 2019) that can be later used for training the network along with labeled examples. But these methods require a moderate number of labeled examples to train a stable GAN model. A recent work (Valvano et al., 2021) trains a stable GAN model with limited annotations of scribbles along with unlabeled data to learn shape priors to obtain high segmentation performance. However, there are some works that use unlabeled examples along with few labeled examples to train either CNN (Zhao et al., 2019) via image registration or conditional GAN model (Chaitanya et al., 2019) optimized to generate synthetic image-label pairs useful for downstream tasks. These methods have shown to perform better than standard augmentations methods.

3. Methods

The objective of this work is to learn discriminative pixel-level representations with intra-class affinity and inter-class separability by leveraging semantic label information that benefits the semantic segmentation task. With this objective, in this section, we describe the proposed semi-supervised method in a limited annotation setting.

We first define the following notation that we use for introducing the proposed method. We define the labeled set with X_L and its image-label pairs with (x_l, y_l). Note that in this paper, we focus on the setting where X_L is limited, e.g. $|X_L| = 1, 2, 8$. Similarly, X_U is the set of pairs $(x_u, \hat{y}_u)$ where x_u denotes unlabeled images and $\hat{y}_u$ denotes the corresponding pseudo-labels obtained and updated using network parameters at certain iterations. The whole dataset is defined as $X = X_L \cup X_U$.

In Fig. 2, we present the network architecture that we use in the proposed method where c_θ is a backbone encoder–decoder network that has two branches: (1) g_ξ is a small network specific for segmentation, (2) h_ϕ is another small network specific for training with contrastive loss.

The proposed semi-supervised segmentation algorithm consists of two optimization steps:

In the first optimization step, we train the networks c_θ and g_ξ by minimizing a supervised segmentation loss L_{seg} using the limited labeled set X_L

$$\hat{\theta}^{(0)}, \hat{\xi}^{(0)} = \arg\min_{\theta,\xi} \frac{1}{|X_L|} \sum_{(x_l, y_l) \in X_L} L_{seg}(y_l, \hat{y}_l) \tag{1}$$

where $\hat{y}_l = g_\xi(c_\theta(x_l))$, we use the dice loss proposed by Milletari et al. (2016) for L_{seg} and (θ, ξ) are initialized randomly. The purpose of

this initial optimization is to obtain reasonable initial pseudo-labels estimates for unlabeled images.

In the second optimization step, we perform joint training of the networks c_θ, $g_{\hat{\xi}}$, and h_ϕ using both supervised segmentation loss L_{seg} on X_L and the proposed local contrastive loss L_{cont} on the whole data X that includes both labeled X_L and unlabeled X_U images.

$$\hat{\theta},\hat{\xi},\hat{\phi} = \arg\min_{\theta,\xi,\phi} \frac{1}{|X_L|} \sum_{(x_l,y_l)\in X_L} L_{seg}(y_l,\hat{y}_l) + \lambda_{cont} \frac{1}{|X|} \sum_{(x,y),(x',y')\in X} L_{cont}((x,y),(x',y')). \tag{2}$$

Note that, (x, y) and (x', y') in L_{cont} are two random samples with replacement from $X = X_L \cup X_U$ which can come from either X_L or X_U. The pseudo-labels of the unlabeled images $x_u \in X_U$ are obtained as $\hat{y}_u = g_{\hat{\xi}^{(t)}}(c_{\hat{\theta}^{(t)}}(x_u))$ using the model parameters $\hat{\xi}^{(t)}$ and $\hat{\theta}^{(t)}$ at iteration t. We initially estimate the pseudo-labels at $t = 0$ and then update them in every P iterations. It is also referred to as *pseudo-labeling* step (refer to Section 4.4 and Section 5.6(b) for more details and analysis, respectively). Also, note that in the second step of optimization, θ and ξ are initialized as $\hat{\theta}^{(0)}$ and $\hat{\xi}^{(0)}$, respectively and ϕ is initialized randomly. λ_{cont} is a weighting parameter for the contrastive loss.

Next, **we introduce L_{cont}, the proposed local pixel-wise contrastive loss**. In the proposed local pixel-wise contrastive loss, we aim to learn discriminative pixel representations based on their class ground-truth labels and pseudo-labels. To achieve this, we optimize to match the pixel representations from the same class label within the image and across different images to be similar while they being dissimilar to the pixel representations that belong to other classes. After the optimization, pixel representations are expected to form a cluster per class as illustrated on the right part of Fig. 2. This allows us to have a better class separation in the feature space by leveraging the contrastive loss on the whole dataset using both ground truth labels of the labeled set and pseudo-label estimates of the unlabeled set.

The feature map after passing an image x through the common network (c_θ) and contrastive branch (h_ϕ) is denoted by $z(x) = h_\phi(c_\theta(x))$ which has dimensions of $H \times W \times D$ where H, W are same as input image dimensions and D is the number of channels. Let us also denote the set of pixel coordinates that belong to a foreground class c for image x as $S_c(x)$ where $1 \le c \le C$ and C is the total number of classes to be segmented. We define the total local contrastive loss between two random samples (x, y) and (x', y') from X as

$$L_{cont}((x,y),(x',y')) = \frac{1}{C}\sum_{c=1}^{C} \frac{1}{|S_c(x)|} \sum_{i\in S_c(x)} L_{i,c}([z(x)]_i, \bar{z}_c(x')) \tag{3}$$

where $[z(x)]_i$ is the feature vector of x at pixel coordinate i, which is D dimensional, and

$$\bar{z}_c(x') = \frac{1}{|S_c(x')|} \sum_{i\in S_c(x')} [z(x')]_i$$

is the mean pixel representation of x' for class c. We define the local contrastive loss between a pixel representation feature vector and a mean pixel representation, possibly coming from a different image, $L_{i,c}(.,.)$ for a class c as follows:

$$L_{i,c}([z(x)]_i, \bar{z}_c(x')) = -\log \frac{e^{\text{sim}([z(x)]_i,\bar{z}_c(x'))/\tau}}{e^{\text{sim}([z(x)]_i,\bar{z}_c(x'))/\tau} + \sum_{k\neq c} e^{\text{sim}([z(x)]_i,\bar{z}_k(x'))/\tau}} \tag{4}$$

where $\text{sim}(a, b) = a^T b/\|a\|\|b\|$ is the cosine similarity to measure the similarity between two representation vectors, and τ denotes the temperature scaling factor as defined in Chen et al. (2020a).

The proposed local pixel-wise contrastive loss defined in Eq. (3) could also be designed to match pixel representation ($[z(x)]_i$) of image x at pixel location i to the pixel representation ($[z(x')]_j$) of image x' at location j. In a similar way, this would encourage similar representations for pixels belonging to the same class c while simultaneously pushing $[z(x)]_i$ away from $[z(x')]_k$ if the corresponding pixels belong to different classes. For a stable and computationally more efficient training, we match $[z(x)]_i$ to the mean representation vector $\bar{z}_c(x')$ and push it away from the mean representation vectors of other classes $\bar{z}_k(x')$.

Two different implementation of the proposed local loss is possible: (1) matching **intra-image** pixel representations: When $(x, y) = (x', y')$ in Eq. (3), the similar class mean representation $\bar{z}_c(x)$ and dissimilar class mean representations $\bar{z}_k(x)$ $(k \neq c)$ are computed from the same image. (2) matching **inter-image** pixel representations: When $(x, y) \neq (x', y')$ in Eq. (3), $\bar{z}_c(x)$ and $\bar{z}_k(x)$ $(k \neq c)$ are computed from different images sampled from X. In the intra-level representation matching, both the similar and dissimilar classes mean representations are computed from within the same image. In the inter-level representation matching, for similar pairs, the mean representation of a class is computed from within the same image as well as from different images in the batch. Similarly, the dissimilar mean representations are computed from within the same image and different images in the batch. With inter-level matching, one can learn more robust representations for the same class structures present in all images in the dataset.

Due to memory limitations, we cannot apply Eq. (3) to pixel representations for all the coordinates $i \in S_c(x)$ in an image since $|S_c(x)|$ can be very large. We deal with this problem by subsampling a smaller set of pixel coordinates $\tilde{S}_c(x) \subset S_c(x)$ in each iteration to fit the GPU memory. Then, we use $\tilde{S}_c(x)$ instead of $S_c(x)$ in Eq. (3). Also, note that the optimization in Eq. (2) can be performed using mini-batch gradient descent by selecting a mini-batch of X_B from X such that X_B contains samples from both X_L and X_U.

4. Experiments

4.1. Datasets

We evaluated the proposed approach on three public MRI datasets.

(a) ACDC Dataset: It contains 100 short-axis MR-cine T1 3D volumes of cardiac anatomy acquired using 1.5T and 3T scanners. The expert annotations are provided for three structures: right ventricle, myocardium, and left ventricle. It was hosted as part of the MICCAI ACDC challenge 2017.

(b) Prostate dataset: It contains 48 T2-weighted MRI 3D volumes of prostate. The expert annotations are provided for two structures of prostate: peripheral zone and central gland. It was hosted in the medical decathlon challenge in MICCAI 2019.

(b) MMWHS dataset: It contains 20 T1 MRI 3D volumes of cardiac anatomy with expert annotations for seven structures: left ventricle, left atrium, myocardium, ascending aorta, pulmonary artery, right ventricle, and right atrium. It was hosted in STACOM and MICCAI 2017 challenges.

4.2. Pre-processing

All the images are bias-corrected using N4 (Tustison et al., 2010) bias correction using the implementation in the ITK toolkit. After this, we apply the below pre-processing steps for all the datasets: (i) each 3D volume (x) is normalized using min–max normalization: $(x - x_1)/(x_{99} - x_1)$ where x_1 and x_{99} denote the 1st and 99th percentile in x. (ii) Next, we re-sample all the 2D images and their corresponding labels into a fixed in-plane resolution r_f using bilinear and nearest neighbor interpolation, respectively. This is followed by cropping or zero padding to a fixed image dimensions of d_f. The in-plane resolution (r_f) and image dimensions (d_f) for the datasets are: (a) ACDC: $r_f = 1.367 \times 1.367$ mm^2 and $d_f = 192 \times 192$, and (b) Prostate: $r_f = 0.6 \times 0.6$ mm^2 and $d_f = 192 \times 192$, and (c) MMWHS: $r_f = 1.5 \times 1.5$ mm^2 and $d_f = 160 \times 160$.

4.3. Network architecture

We use an UNet (Ronneberger et al., 2015) based architecture that consists of a common encoder and decoder networks denoted by c_θ. From the last layer of the decoder, we have 2 different branches of smaller networks: one for segmentation (g_ξ) and one for contrastive learning (h_ϕ). The encoder consists of 6 convolutional blocks, each block consists of two 3×3 convolutions followed by a 2×2 max pool layer with a stride of 2. The decoder consists of 5 convolutional blocks, each block consists of an upsampling layer with a factor of 2, followed by concatenation from a corresponding layer of the encoder via a skip connection, that is followed by two 3×3 convolutions. The segmentation task network (g_ξ) consists of three 3×3 convolutions, followed by a softmax layer to output the segmentation mask, used in the dice loss (Milletari et al., 2016). The contrastive task network (h_ϕ) consists of two 1×1 convolution layers that outputs feature maps f of dimensions $H \times W \times D$ that are used for local contrastive loss computation. We use 1×1 convolution to mimic the behavior of dense layers in the projection head in learning global level representations in the popular contrastive learning work (Chen et al., 2020a). All the layers except the last layers of both segmentation task and contrastive loss computation have batch normalization (Ioffe and Szegedy, 2015) and ReLU activation layers.

4.4. Experimental setup

(i) For the proposed method, for initial 5000 iterations, we train with only the segmentation loss using the labeled set $X_L = (x_l, y_l)$. In this step, only the common encoder–decoder network parameters (θ) and segmentation-specific layers parameters (ξ) are updated. (ii) Next, we estimate the pseudo-labels $\hat{y_u} \in Y_U$ of unlabeled set X_U using this trained network. (iii) After this, we perform the optimization given in Eq. (2) for P=5000 iteration. Here, all the network parameters are updated: encoder–decoder network (θ), segmentation specific layers (ξ), and contrastive loss specific layers (ϕ). (iv) We iterate the process of pseudo-label initial estimation/re-estimation (ii) and joint loss training (iii) for a total of 3 pseudo-labeling steps in this work resulting in total no of iterations to be 15000 for all the datasets. This value of 3 was found to be a good solution in the compared self-training work (Bai et al., 2017).

4.5. Training details

The network architecture and hyper-parameters such as number of layers in encoder–decoder, kernel size and others where chosen based on prior work (Baumgartner et al., 2017) that explored different combinations of these hyper-parameters. As we mentioned in Section 3, the proposed method has two optimization steps. We perform both optimizations using Adam optimizer (Kingma and Ba, 2015) with a learning rate of 10^{-3}. We set the batch size $|X_B| = 20$ in both optimization steps. During the first step we solve the problem given in Eq. (1), i.e., we sample the images in X_B only from the limited labeled set X_L and optimize the supervised segmentation loss. We run this optimization for 5000 iterations. During the second step we solve the problem given in Eq. (2). In each batch, we sample 10 image-label pairs from X_L and 10 image-pseudo-label pairs from X_U. We run this optimization for 15000 iterations where we initially estimate the pseudo-labels at iteration 0 and update them every P=5000 iterations.

For the local contrastive loss, we apply random intensity transformations (contrast plus brightness) on the input image batch used to compute the contrastive loss. For the segmentation loss, we choose dice loss (Milletari et al., 2016) and apply random data augmentations such as crop, scale, rotation, flip, random elastic deformations, and random intensity transformations of contrast and brightness on the batch of labeled images.

For the contrastive loss, we sample $|\tilde{S}_c(x)|$ number of positive pixel representations for each class c and each image x. The default value of $|\tilde{S}_c(x)|$ is set as 3 for all the experiments unless stated otherwise. The number of classes C (structures of interest) is different for each dataset, whose value is 3 for ACDC, 2 for Prostate, and 7 for MMWHS. The temperature parameter τ chosen for contrastive loss is 0.1 as defined in Chaitanya et al. (2020) for the medical datasets. The λ_{cont} value as mentioned in Eq. (2) is set to 0.1 in all experiments unless otherwise stated. We use the number of channels as 16 in the last layer of the contrastive network branch h_ϕ. Therefore, the pixel representations are 16-dimensional for all the datasets. We used the highest Dice score model on the validation set during training to select the model for the final evaluation on the test set.

Dataset Split: We use the following size of unlabeled (X_U) and test sets (X_{ts}): (a) $|X_U|$=52, $|X_{ts}|$=20 for ACDC, (b) $|X_U|$=22, $|X_{ts}|$=15 for Prostate, and (c) $|X_U|$=10, $|X_{ts}|$=10 for MMWHS. For the limited annotation setting evaluation, we use a small labeled (X_L) and validation sets (X_{val}). The validation set ($|X_{val}|$) size is always fixed to two 3D volumes for all datasets. For the labeled set, we experiment with three different sizes of $|X_L|$ = 1, 2, and 8 3D volumes. The earlier works use two major ways to define these labeled training set sizes: (a) absolute numbers (Bai et al., 2019; Chaitanya et al., 2019, 2020; Hooper et al., 2020; Chen et al., 2020b; Ouyang et al., 2020; Peng et al., 2021b; Ouyang et al., 2022) and (b) percentage or fractions of datasets (Luo et al., 2021; Chen et al., 2021; Verma et al., 2019). In this work, we prefer absolute numbers for the following two reasons. First, from the application perspective, absolute numbers allows realistic assessment of the proposed technique. Second, with absolute numbers it is easier to compare different segmentation tasks as using absolute numbers allows results on data sets with largely different number of samples, while percentages prohibits such comparisons.

Evaluation: The segmentation performance is evaluated using Dice's similarity coefficient (DSC) (Dice, 1945). For all the experiments, we report the test set's $|X_{ts}|$ mean DSC computed over all the structures without the background over 6 different runs. For each run, we construct different sets of X_L and X_{val} by randomly sampling from the available set of volumes.

5. Results and discussion

In Table 1, we present the results of baseline, existing semi-supervised methods in the literature, the proposed method, and concurrent contrastive learning works.

5.1. Baseline and semi-supervised methods

From Table 1, we can observe that semi-supervised methods yield a significant boost in performance for limited labeled settings of $|X_L|$=1 or 2 3D volumes for all the datasets compared to the baseline. We observe that methods like self-training and noisy student with $|X_L|$=2 close the gap to benchmark that is trained with a large number of annotations to less than 0.1 DSC for Prostate and MMWHS datasets, and 0.2 DSC for ACDC dataset.

5.2. Proposed method

We observe that the proposed method yields better results than the baseline and compared semi-supervised methods from Table 1. We present the qualitative results in Fig. 3 where we can observe the improvements in the predicted masks for finer or harder to segment structures for all the datasets. Results suggest that using pseudo-labels only in a contrastive loss, as done in the proposed method, yields higher performance gains compared to using them in the segmentation loss, as is often done in popular self-training methods (Bai et al., 2017; Xie et al., 2020a) and consistency regularization methods like DTC (Luo et al., 2021), CPS (Chen et al., 2021), ICT (Verma et al.,

Table 1
Comparison of the proposed method with other semi-supervised learning, data augmentation methods, and concurrent contrastive learning works. The proposed pseudo-label joint training provides better results than compared methods for all datasets and most $|X_{tr}|$ values. In each column, the best values among all methods are underlined. $*$ indicates the statistically significant improvements between the proposed method compared to the best semi-supervised method and concurrent contrastive learning method using Wilcoxon's paired t-test with a threshold p-value of 0.01.

Method	ACDC			Prostate			MMWHS		
	$\|X_L\|$=1	$\|X_L\|$=2	$\|X_L\|$=8	$\|X_L\|$=1	$\|X_L\|$=2	$\|X_L\|$=8	$\|X_L\|$=1	$\|X_L\|$=2	$\|X_L\|$=8
Baseline									
Random init.	0.614	0.702	0.844	0.489	0.550	0.636	0.451	0.637	0.787
Semi-supervised Methods									
Self-training (Bai et al., 2017)	0.690	0.749	0.860	0.551	0.598	0.680	0.563	0.691	0.801
Mixup (Zhang et al., 2017a)	0.695	0.785	0.863	0.543	0.593	0.661	0.561	0.690	0.796
Data Augment (Chaitanya et al., 2019)	0.731	0.786	0.865	0.585	0.597	0.667	0.529	0.661	0.785
Adversarial Tr. Zhang et al. (2017b)	0.536	0.654	0.791	0.487	0.544	0.586	0.482	0.655	0.779
DTC (Luo et al., 2021)	0.597	0.687	0.849	0.511	0.587	0.658	0.432	0.610	0.791
CPS (Chen et al., 2021)	0.666	0.726	0.846	0.504	0.567	0.630	0.567	0.661	0.791
ICT (Verma et al., 2019)	0.638	0.732	0.854	0.544	0.567	0.677	0.518	0.639	0.797
Noisy Student (Xie et al., 2020a)	0.632	0.737	0.836	0.556	0.601	0.668	0.593	0.685	0.780
Concurrent Contrastive Methods									
Two step Tr. Zhao et al. (2020) (fully-sup)	0.568	0.717	0.841	0.561	0.596	0.686	0.509	0.661	0.795
Two step Tr. Zhao et al. (2020) (semi-sup)	0.622	0.747	0.843	0.568	0.606	0.688	0.559	0.704	0.804
Joint Tr. with pseudo-labels in seg. loss (Alonso et al., 2021)	0.420	0.603	0.749	0.435	0.495	0.545	0.470	0.510	0.638
Proposed Method (Ours)									
Pseudo-labels joint Tr. (intra)	0.761*	0.845*	0.881*	0.571*	0.613*	0.693	0.599*	0.721*	0.803
Pseudo-labels joint Tr. (inter)	0.759*	0.831*	0.883*	0.578*	0.618*	0.696	0.572*	0.719*	0.811
Benchmark									
Training with large $\|X_L\|$	($\|X_L\|$ = 78) 0.912			($\|X_L\|$ = 20) 0.697			($\|X_L\|$ = 8) 0.787		

2019) in medical image analysis. DTC (Luo et al., 2021), CPS (Chen et al., 2021), ICT (Verma et al., 2019) and MC-Net (Wu et al., 2021) have low performance for limited labeled case of $|X_L|$=1 or 2 3D volumes compared to older self-training methods (Bai et al., 2017; Xie et al., 2020a) while they perform better with a large labeled setting of $|X_L|$=8. This low performance for these latest methods contrary to the results reported in their original works can be due to: (a) the lowest evaluated labeled settings were $|X_L|$=16 in DTC (Luo et al., 2021) and $|X_L|$=8 or 16 in MC-Net (Wu et al., 2021) which is higher than our lowest evaluation of $|X_L|$=1 or 2, and (b) most of them only perform left atrium segmentation while we segment multiple structures and (c) use a different cardiac dataset to ours for evaluation. This can be due to initial pseudo-label estimates being erroneous and thus deteriorating the performance if directly used for the segmentation loss optimization, especially for the limited labeled case of $|X_L|$=1 or 2 3D volumes. We hypothesize that using pseudo-labels only in the contrastive loss and not in the segmentation loss acts as regularization, and prevent erroneous pseudo-label mask information propagating into the segmentation task-specific layers. We believe that this provides us with the additional improvements that could have hindered the self-training approach. Also, another reason for improvements can be due to learning a stronger pixel-level representation by matching representations of the same class to be similar across images with the defined contrastive loss, which typical segmentation loss of cross-entropy or dice loss cannot learn. Also, we see that the re-estimation of pseudo-labels for the unlabeled data iteratively in the training leads to continuous improvements through the training as shown in Fig. 5 (more discussion in Section 5.5).

We evaluated two pixel representation matching schemes: intra-image and inter-image pixel representation matching. Both matching schemes yield improvements over compared methods. Inter-image pixel representation matching performs better than intra-image pixel representation matching in 5 out of 9 cases. This can be because matching pixel representations across images from different volumes from the same class can provide the network additional cues over intra-image matching to learn more robust representations for each semantic class

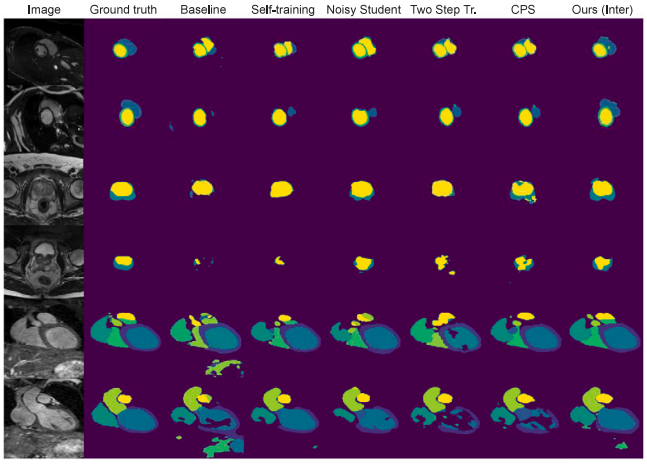

Fig. 3. We present the visual results on two random test images from each dataset for the proposed method and compared semi-supervised methods. We observe improvement in predicted segmentation masks more prominently for the finer or harder to segment structures.

in the dataset. For a limited labeled case of X_L=1, the intra-image matching performs better on ACDC and MMWHS datasets than inter-image as the quality of the pseudo-labels can be of low quality in this case, and thereby, matching erroneous class pixel representations across images can hinder the performance gains that can be obtained.

5.3. Concurrent contrastive semi-supervised methods

In the concurrent works (Alonso et al., 2021; Zhao et al., 2020), a different set of semi-supervised contrastive learning frameworks are proposed. In Zhao et al. (2020), a two-step training is proposed that consists of pre-training with contrastive loss and later fine-tuning to

Table 2
We compare baseline, self-training, and proposed method trained over a pre-trained network initialization (pre-trained init.) instead of random initialization (random init.). This initialization was learnt via contrastive loss based pre-training as done in Chaitanya et al. (2020). We see that baseline, self-training, and the proposed method all benefit from better initialization and lead to higher gains. We can conclude that the better pre-trained initialization and proposed method are complementary.

Method	ACDC			Prostate			MMWHS		
	$\|X_L\|$=1	$\|X_L\|$=2	$\|X_L\|$=8	$\|X_L\|$=1	$\|X_L\|$=2	$\|X_L\|$=8	$\|X_L\|$=1	$\|X_L\|$=2	$\|X_L\|$=8
Baseline with random init.	0.614	0.702	0.844	0.489	0.550	0.636	0.451	0.637	0.787
Baseline with pre-trained init. (Chaitanya et al., 2020)	0.725	0.789	0.872	0.579	0.619	0.684	0.569	0.694	0.794
Self-training((Bai et al., 2017)) with pre-trained init. (Chaitanya et al., 2020)	0.745	0.802	0.881	0.607	0.634	0.698	0.647	0.727	0.806
Proposed method (inter) with pre-trained init. (Chaitanya et al., 2020)	0.774	0.845	0.887	0.612	0.641	0.692	0.651	0.734	0.814
Proposed method (inter) with random init.	0.759	0.831	0.883	0.578	0.618	0.696	0.572	0.719	0.811

final task with labeled set, which is a fully supervised setting. This two-step training strategy can also be used in a semi-supervised setting by using both the labeled annotations and pseudo-labels of the unlabeled set for both stages of the training. We perform experiments in both settings and present these results in the initial two rows of "concurrent contrastive methods" section in Table 1 in fully supervised and semi-supervised settings, respectively. With the results presented, we can infer that joint training can be more beneficial than disjoint training, and using pseudo-labels in segmentation loss can hinder the maximum possible gains that could be obtained, especially, in a limited labeled setting of $|X_L|$= 1 and 2. For Alonso et al. (2021), they perform joint training where both contrastive and segmentation losses are computed using both labeled and unlabeled sets. They devise student–teacher networks, use additional memory bank to store class representations and use entropy minimization. We rather propose a simpler solution without using an additional teacher network, memory banks and entropy minimization. For an equivalent comparison with the proposed approach, we evaluate this joint training approach with only one network, and without additional components. Here, both the labeled annotations and pseudo-labels of the unlabeled set are used in both the contrastive and segmentation losses. We present these results in the final row of "concurrent contrastive methods" section in Table 1, where we see that using pseudo-labels in segmentation loss computation even in joint training can lead to lower gains similar to our observation in the self-training approach. Overall, we obtain higher gains than these methods for all settings across all the datasets where the gains are more prominent for the ACDC dataset.

5.4. Better model initialization from contrastive loss pre-training

Recent work by Chaitanya et al. (2020) applies a contrastive learning framework to learn a good initialization for medical image segmentation tasks and achieve state-of-the-art results and perform better than random initialization. We use this pre-trained network initialization for training the baseline, self-training, and proposed method whose results are presented in Table 2. Firstly, we obtain larger improvements when fine-tuning a baseline model with this pre-trained initialization over random initialization as shown in the first two rows. Secondly, we get a higher mean DSC for both self-training (row 3) and the proposed method (row 4) with this initialization compared to the random initialization (row 1). Also, the proposed method does better with this pre-trained initialization (row 4) compared to random initialization (row 5) because we have a better initial baseline model that yields more accurate initial pseudo-label estimates compared to the random initialization. Thirdly, for this initialization as well, we get higher gains with the proposed method over self-training where the gains on the ACDC dataset are more significant while the gains are smaller on the remaining two datasets. Lastly, we see that proposed joint training is complementary to a good network initialization learned via pre-training.

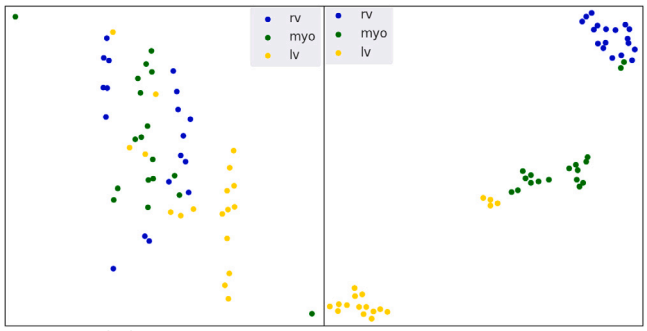

Fig. 4. TSNE plot of pixel representations of the last layer of the common network (c_θ for three structures of ACDC (right ventricle (rv), myocardium (myo), left ventricle (lv)) for some randomly sampled unlabeled images for the methods: (a) self-training and (b) proposed method with local contrastive loss. For the proposed approach, we observe a better intra-class affinity and inter-class separation between pixel representations compared to self-training, and thereby leading to higher segmentation performance.

5.5. Visualization of results of the proposed method

In Fig. 4, we present a tSNE plot of the pixel representations for the three cardiac structures of the right ventricle (rv), myocardium (myo), left ventricle (lv) of the ACDC dataset at the last layer of the common network (output of c_θ) for both the self-training model and proposed method model. In the proposed method, we observe that the pixel representations of the three structures have more compact clusters for each class (intra-class affinity), and different class representations are more separable (inter-class separability) compared to the pixel representations from the self-training model. With the contrastive loss, the network learns a better intra-class affinity and inter-class separability compared to a segmentation loss that may not be optimal to learn such semantic relationships between pixels.

In Fig. 5, for the setting of $|X_L|$=1 3D volume on ACDC dataset, we present the test set's mean DSC for three runs where a different set of training volume and validation volumes chosen for self-training and proposed method. The initial model was trained using supervised segmentation loss using only a limited labeled set of X_L=1. Later, we use this model to estimate pseudo-labels such that both methods start with same pseudo-label estimates. After this, we perform the respective iterative training as proposed in self-training and proposed approach with three pseudo-labeling steps with pseudo-labels updated every 5000 iterations. Here, we observe that the test set's mean DSC improves after the three pseudo-labeling steps for both self-training and proposed method. The improvements are larger for the proposed method from the first pseudo-labeling step compared to self-training across all runs. Also, we see that the self-training approach's performance gains become constant from second pseudo-labeling step while the proposed approach yields further improvements at the second step. After the third update step, even proposed method's gains become constant across all 3 runs.

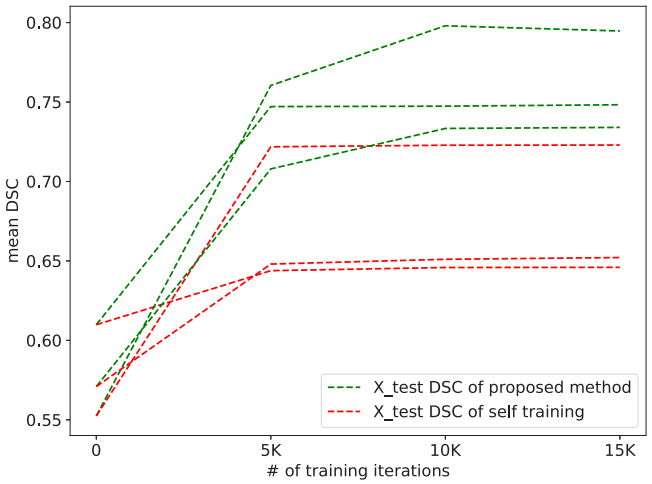

Fig. 5. Test set mean DSC reported on ACDC dataset over the 3 pseudo-labeling steps for self-training and proposed method, where each step denotes 5000 iterations of training. The proposed approach yields consistent and higher performance gains for first and second pseudo-labeling steps compared to self-training that do not provide any additional gains from second update step.

5.6. Hyper-parameter analysis

Since a large number of computational resources are required for each hyper-parameter experiment on each dataset, we evaluate different hyper-parameters on ACDC dataset only.

(a) λ_{cont} for contrastive loss ($\mathbf{L}_{cont}$): In Table 3, we present the results for different values of lambda (λ_{cont}) as defined in Eq. (2) to dictate the contribution of the contrastive loss in the total loss used for updating the network weights. We observe that the performance deteriorates when we use a very small lambda value for the contrastive loss contribution when the number of labeled examples is low. This coefficient does not have a distinguishable effect when $|X_L| = 8$.

(b) Number of pseudo-labeling steps: Here, we analyze the effect of number of pseudo-labeling steps during the training. We define the total number of training iterations to be 15000 for all the below experiments. We evaluate three values of pseudo-labeling steps of 2, 3, and 4 where the pseudo-labels are updated in every P iterations of P=7500 (15000/2), P=5000 (15000/3), and P=3750 (15000/4), respectively. The results are presented in Table 4. We observe for both intra and inter-representation matching schemes that the performance difference between the evaluated values of number of pseudo-labeling steps are small indicating the stable behavior of training and gains in performance obtained with the proposed approach.

(c) Number of positive pixel representations $|S_c(x)|$ sampled per class c per image x: Here, the number of positive pixel representations $|S_c(x)|$ sampled per class per image to match its class mean representation are varied between the values of 3, 5, and 10. In Table 5, for both intra- and inter-image representation matching schemes, we do not see a discernible pattern in performance with respect to the number of pixels sampled. This shows that the model is stable with respect to the final performance against changes in the number of positive pixel representations sampled per class per image.

(d) Improving the quality of the pseudo-labels of unlabeled set: Here, we control the quality of the pseudo-label predictions of unlabeled volumes used in the contrastive loss, and the results are presented in Table 6. The threshold value defined denotes the DSC overlap between two estimated masks of the same volume under two different sets of transformations like rotation, translation, scaling, etc. This idea is borrowed from the consistency regularization works (Laine and Aila, 2016; Bortsova et al., 2019; Li et al., 2020a) in the literature to ensure that the pseudo-label predictions of the unlabeled set are accurate and do not vary drastically under transformations. The value of 0.9 denotes having a highly confident estimate where the two estimated masks of the same volume under different transformations are very similar and have an overlap dice of 0.9 while 0.7 denotes a less confident estimate where the two estimated masks are less similar to each other. The value of 0 denotes that no consistency regularization was applied, and all the pseudo-label predictions of all volumes were used in the contrastive loss. We observe that using highly confident unlabeled set predictions does not necessarily yield higher gains compared to using all the predictions of the unlabeled set. At threshold value of 0, we also observe that inter-image representation matching performs worse than intra-image representation matching for lower number of labeled examples of $|X_L|$=1 or 2. This can be due to the poor quality of pseudo-labels estimated, which when used to match the representations of same class across subjects leads to lower gains.

Table 3
Fine-tuning with different lambda (λ_{cont}) values to control the influence of contrastive loss in the proposed method with a random network initialization on ACDC dataset. We observe the proposed method to be relatively stable for λ_{cont} values of 0.1, 1 and the performance drops slightly for low value of 0.01.

Representation matching scheme	λ_{cont} value	$\lvert X_L\rvert$=1	$\lvert X_L\rvert$=2	$\lvert X_L\rvert$=8
Intra	1	0.763	0.817	0.886
Intra	0.1	0.761	0.845	0.881
Intra	0.01	0.731	0.801	0.890
Inter	1	0.749	0.849	0.898
Inter	0.1	0.759	0.831	0.883
Inter	0.01	0.722	0.815	0.888

Table 4
We evaluate the effect on segmentation performance when we vary the pseudo-labeling steps on ACDC dataset. We do not observe significant difference between different values of pseudo-labeling steps.

Representation matching scheme	No. of pseudo-labeling steps	$\lvert X_L\rvert$=1	$\lvert X_L\rvert$=2	$\lvert X_L\rvert$=8
Intra	2	0.764	0.836	0.878
Intra	3	0.761	0.845	0.881
Intra	4	0.771	0.813	0.884
Inter	2	0.728	0.839	0.895
Inter	3	0.759	0.831	0.883
Inter	4	0.777	0.83	0.888

Table 5
Fine-tuning with different number of positive representation examples $|S_c(x)|$ per class c and per image x for the proposed method with a random network initialization on ACDC dataset. We do not see any major differences in performance denoting the stable training.

Representation matching scheme	$\lvert S_c(x)\rvert$	$\lvert X_L\rvert$=1	$\lvert X_L\rvert$=2	$\lvert X_L\rvert$=8
Intra	3	0.761	0.845	0.881
Intra	5	0.772	0.847	0.882
Intra	10	0.752	0.818	0.885
Inter	3	0.759	0.831	0.883
Inter	5	0.750	0.847	0.882
Inter	10	0.732	0.862	0.897

6. Conclusion

It is challenging to deploy deep learning-based models with high performance for medical image segmentation due to the requirement of a large set of annotations. To alleviate this, we propose a semi-supervised method that uses many unlabeled images and a limited

Table 6
Fine-tuning with different threshold values to control the quality of pseudo-labels of the unlabeled volumes used in the contrastive loss computation of the proposed method with a random network initialization on ACDC dataset. We see that selecting higher quality of pseudo-labels with threshold values of 0.9 and 0.8 does not lead to more gains over using all the pseudo-labels (threshold value=0).

Representation matching scheme	threshold value	$\|X_L\|$=1	$\|X_L\|$=2	$\|X_L\|$=8
Intra	0	0.761	0.845	0.881
Intra	0.9	0.748	0.841	0.890
Intra	0.8	0.762	0.839	0.888
Intra	0.7	0.769	0.856	0.878
Inter	0	0.759	0.831	0.883
Inter	0.9	0.779	0.832	0.891
Inter	0.8	0.791	0.860	0.882
Inter	0.7	0.732	0.847	0.889

set of annotations to yield high segmentation performance. We introduce a joint training framework where a pixel-wise contrastive loss is defined over pseudo-labels of unlabeled images and limited labeled images with the traditional segmentation loss applied on only the labeled set. We perform iterative training to improve the quality of the pseudo-labels. With the proposed contrastive loss, we learn better intra-class compactness and inter-class separability for the segmented classes in the dataset compared to typical segmentation loss-based methods like self-training. We perform an extensive evaluation of the proposed method on three MRI datasets and obtain high segmentation performance in limited annotation settings. We get higher performance gains than the compared state-of-the-art semi-supervised methods and concurrent contrastive learning methods. We also show that a good network initialization learned via pre-training with unlabeled images is complementary to the proposed method and can be combined to obtain higher gains. Also, using only the labeled set in the segmentation loss compared to earlier approaches is one of the essential detail in the performance gains obtained in the proposed method.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The code is made publicly available on github whose link is mentioned in the manuscript and dataset used are public datasets

Acknowledgments

The presented work was partly funding by: 1. Clinical Research Priority Program (CRPP) Grant on Artificial Intelligence in Oncological Imaging Network, University of Zurich, 2. Swiss Platform for Advanced Scientific Computing (PASC), coordinated by Swiss National Super-computing Centre (CSCS), 3. Personalized Health and Related Technologies (PHRT), project number 222, ETH domain, 4. University Hospital Zurich, Switzerland. We also thank Nvidia for their GPU donation.

References

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S., 2012. SLIC superpixels compared to state-of-the-art superpixel methods. IEEE Trans. Pattern Anal. Mach. Intell. 34 (11), 2274–2282.

Alonso, I., Sabater, A., Ferstl, D., Montesano, L., Murillo, A.C., 2021. Semi-supervised semantic segmentation with pixel-level contrastive learning from a class-wise memory bank. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8219–8228.

Arbelaez, P., Maire, M., Fowlkes, C., Malik, J., 2010. Contour detection and hierarchical image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 33 (5), 898–916.

Arbeláez, P., Pont-Tuset, J., Barron, J.T., Marques, F., Malik, J., 2014. Multiscale combinatorial grouping. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 328–335.

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. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 541–549.

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. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 253–260.

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. In: International Workshop on Statistical Atlases and Computational Models of the Heart. Springer, pp. 111–119.

Berthelot, D., Carlini, N., Cubuk, E.D., Kurakin, A., Sohn, K., Zhang, H., Raffel, C., 2019a. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. arXiv preprint arXiv:1911.09785.

Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., Raffel, C., 2019b. Mixmatch: A holistic approach to semi-supervised learning. arXiv preprint arXiv:1905.02249.

Blundell, C., Cornebise, J., Kavukcuoglu, K., Wierstra, D., 2015. Weight uncertainty in neural network. In: International Conference on Machine Learning. PMLR, pp. 1613–1622.

Bortsova, G., Dubost, F., Hogeweg, L., Katramados, I., de Bruijne, M., 2019. Semi-supervised medical image segmentation via learning consistency under transformations. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 810–818.

Bowles, C., Chen, L., Guerrero, R., Bentley, P., Gunn, R., Hammers, A., Dickie, D.A., Hernández, M.V., Wardlaw, J., Rueckert, D., 2018. GAN augmentation: Augmenting training data using generative adversarial networks. arXiv preprint arXiv:1810.10863.

Cai, J., Zhang, Z., Cui, L., Zheng, Y., Yang, L., 2019. Towards cross-modal organ translation and segmentation: a cycle-and shape-consistent generative adversarial network. Med. Image Anal. 52, 174–184.

Camarasa, R., Bos, D., Hendrikse, J., Nederkoorn, P., Kooi, E., van der Lugt, A., de Bruijne, M., 2020. Quantitative comparison of Monte-Carlo dropout uncertainty measures for multi-class segmentation. In: Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Graphs in Biomedical Image Analysis. Springer, pp. 32–41.

Cao, X., Chen, H., Li, Y., Peng, Y., Wang, S., Cheng, L., 2020. Uncertainty aware temporal-ensembling model for semi-supervised abus mass segmentation. IEEE Trans. Med. Imaging 40 (1), 431–443.

Chaitanya, K., Erdil, E., Karani, N., Konukoglu, E., 2020. Contrastive learning of global and local features for medical image segmentation with limited annotations. Adv. Neural Inf. Process. Syst. 33.

Chaitanya, K., Karani, N., Baumgartner, C.F., Becker, A., Donati, O., Konukoglu, E., 2019. Semi-supervised and task-driven data augmentation. In: Chung, A.C.S., Gee, J.C., Yushkevich, P.A., Bao, S. (Eds.), Information Processing in Medical Imaging. Springer International Publishing, Cham, pp. 29–41.

Chapelle, O., Scholkopf, B., Zien, A., 2009. Semi-supervised learning (Chapelle, O. et al., Eds.; 2006)[book reviews]. IEEE Trans. Neural Netw. 20 (3), 542.

Chen, L., Bentley, P., Mori, K., Misawa, K., Fujiwara, M., Rueckert, D., 2019. Self-supervised learning for medical image analysis using image context restoration. Med. Image Anal. 58, 101539.

Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020a. A simple framework for contrastive learning of visual representations. arXiv preprint arXiv:2002.05709.

Chen, C., Qin, C., Qiu, H., Ouyang, C., Wang, S., Chen, L., Tarroni, G., Bai, W., Rueckert, D., 2020b. Realistic adversarial data augmentation for MR image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 667–677.

Chen, X., Yuan, Y., Zeng, G., Wang, J., 2021. Semi-supervised semantic segmentation with cross pseudo supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2613–2622.

Cireşan, D.C., Meier, U., Masci, J., Gambardella, L.M., Schmidhuber, J., 2011. High-performance neural networks for visual object classification. arXiv preprint arXiv:1102.0183.

Costa, P., Galdran, A., Meyer, M.I., Niemeijer, M., Abràmoff, M., Mendonça, A.M., Campilho, A., 2018. End-to-end adversarial retinal image synthesis. IEEE Trans. Med. Imaging 37 (3), 781–791.

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. In: International Conference on Information Processing in Medical Imaging. Springer, pp. 554–565.

Dice, L.R., 1945. Measures of the amount of ecologic association between species. Ecology 26 (3), 297–302.

Doersch, C., Gupta, A., Efros, A.A., 2015. Unsupervised visual representation learning by context prediction. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1422–1430.

Dosovitskiy, A., Springenberg, J.T., Riedmiller, M., Brox, T., 2014. Discriminative unsupervised feature learning with convolutional neural networks. In: Advances in Neural Information Processing Systems. pp. 766–774.

Eaton-Rosen, Z., Bragman, F., Ourselin, S., Cardoso, M.J., 2018. Improving data augmentation for medical image segmentation. In: International Conference on Medical Imaging with Deep Learning.

Fan, D.P., Zhou, T., Ji, G.P., Zhou, Y., Chen, G., Fu, H., Shen, J., Shao, L., 2020. Inf-net: Automatic covid-19 lung infection segmentation from ct images. IEEE Trans. Med. Imaging 39 (8), 2626–2637.

Fang, K., Li, W.J., 2020. Dmnet: Difference minimization network for semi-supervised segmentation in medical images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 532–541.

Felzenszwalb, P.F., Huttenlocher, D.P., 2004. Efficient graph-based image segmentation. Int. J. Comput. Vis. 59 (2), 167–181.

Fotedar, G., Tajbakhsh, N., Ananth, S., Ding, X., 2020. Extreme consistency: Overcoming annotation scarcity and domain shifts. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 699–709.

Gal, Y., Ghahramani, Z., 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: International Conference on Machine Learning. PMLR, pp. 1050–1059.

Gidaris, S., Singh, P., Komodakis, N., 2018. Unsupervised representation learning by predicting image rotations. arXiv preprint arXiv:1803.07728.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets. In: Advances in Neural Information Processing Systems. pp. 2672–2680.

Graham, S., Chen, H., Gamper, J., Dou, Q., Heng, P.A., Snead, D., Tsang, Y.W., Rajpoot, N., 2019. MILD-Net: Minimal information loss dilated network for gland instance segmentation in colon histology images. Med. Image Anal. 52, 199–211.

Grandvalet, Y., Bengio, Y., 2005. Semi-supervised learning by entropy minimization. In: Advances in Neural Information Processing Systems. pp. 529–536.

Gutmann, M., Hyvärinen, A., 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. pp. 297–304.

Hadsell, R., Chopra, S., LeCun, Y., 2006. Dimensionality reduction by learning an invariant mapping. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 2. CVPR'06, IEEE, pp. 1735–1742.

Han, L., Huang, Y., Dou, H., Wang, S., Ahamad, S., Luo, H., Liu, Q., Fan, J., Zhang, J., 2020. Semi-supervised segmentation of lesion from breast ultrasound images with attentional generative adversarial network. Comput. Methods Programs Biomed. 189, 105275.

He, K., Fan, H., Wu, Y., Xie, S., Girshick, R., 2019. Momentum contrast for unsupervised visual representation learning. arXiv preprint arXiv:1911.05722.

Hénaff, O.J., Koppula, S., Alayrac, J.B., Oord, A.v.d., Vinyals, O., Carreira, J., 2021. Efficient visual pretraining with contrastive detection. arXiv preprint arXiv:2103.10957.

Hénaff, O.J., Razavi, A., Doersch, C., Eslami, S., Oord, A.v.d., 2019. Data-efficient image recognition with contrastive predictive coding. arXiv preprint arXiv:1905.09272.

Hjelm, D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., Bengio, Y., 2019. Learning deep representations by mutual information estimation and maximization. In: ICLR 2019. ICLR.

Hong, J., Park, B.y., Park, H., 2017. Convolutional neural network classifier for distinguishing barrett's esophagus and neoplasia endomicroscopy images. In: 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. EMBC, IEEE, pp. 2892–2895.

Hooper, S., Wornow, M., Seah, Y.H., Kellman, P., Xue, H., Sala, F., Langlotz, C., Re, C., 2020. Cut out the annotator, keep the cutout: better segmentation with weak supervision. In: International Conference on Learning Representations.

Hu, X., Zeng, D., Xu, X., Shi, Y., 2021. Semi-supervised contrastive learning for label-efficient medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 481–490.

Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: ICML. pp. 448–456.

Jaus, A., Yang, K., Stiefelhagen, R., 2021. Panoramic panoptic segmentation: Towards complete surrounding understanding via unsupervised contrastive learning. arXiv preprint arXiv:2103.00868.

Jungo, A., Reyes, M., 2019. Assessing reliability and challenges of uncertainty estimations for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 48–56.

Kamnitsas, K., Ferrante, E., Parisot, S., Ledig, C., Nori, A.V., Criminisi, A., Rueckert, D., Glocker, B., 2016. DeepMedic for brain tumor segmentation. In: International Workshop on Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. Springer, pp. 138–149.

Kamnitsas, K., Ledig, C., Newcombe, V.F., Simpson, J.P., Kane, A.D., Menon, D.K., Rueckert, D., Glocker, B., 2017. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Med. Image Anal. 36, 61–78.

Kendall, A., Gal, Y., 2017. What uncertainties do we need in bayesian deep learning for computer vision? arXiv preprint arXiv:1703.04977.

Kingma, D.P., Ba, J., 2015. Adam: A method for stochastic optimization. In: ICLR.

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

Lee, H.H., Tang, Y., Yang, Q., Yu, X., Bao, S., Landman, B.A., Huo, Y., 2021. Attention-guided supervised contrastive learning for semantic segmentation. arXiv preprint arXiv:2106.01596.

Lee, D.H., et al., 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: Workshop on Challenges in Representation Learning, Vol. 3, No. 2. ICML.

Li, X., Yu, L., Chen, H., Fu, C.W., Xing, L., Heng, P.A., 2020a. Transformation-consistent self-ensembling model for semisupervised medical image segmentation. IEEE Trans. Neural Netw. Learn. Syst..

Li, S., Zhang, C., He, X., 2020b. Shape-aware semi-supervised 3d semantic segmentation for medical images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 552–561.

Luo, X., Chen, J., Song, T., Wang, G., 2021. Semi-supervised medical image segmentation through dual-task consistency. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, No. 10. pp. 8801–8809.

Mehrtash, A., Wells, W.M., Tempany, C.M., Abolmaesumi, P., Kapur, T., 2020. Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. IEEE Trans. Med. Imaging 39 (12), 3868–3878.

Milletari, F., Navab, N., Ahmadi, S.A., 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision. 3DV, IEEE, pp. 565–571.

Misra, I., van der Maaten, L., 2019. Self-supervised learning of pretext-invariant representations. arXiv preprint arXiv:1912.01991.

Miyato, T., Maeda, S.i., Koyama, M., Ishii, S., 2018. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. IEEE Trans. Pattern Anal. Mach. Intell. 41 (8), 1979–1993.

Nair, T., Precup, D., Arnold, D.L., Arbel, T., 2020. Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. Med. Image Anal. 59, 101557.

Nguyen, D.T., Dax, M., Mummadi, C.K., Ngo, T.P.N., Nguyen, T.H.P., Lou, Z., Brox, T., 2019. Deepusps: Deep robust unsupervised saliency prediction with self-supervision. Adv. Neural Inf. Process. Syst. 32.

Nie, D., Gao, Y., Wang, L., Shen, D., 2018. ASDNet: attention based semi-supervised deep networks for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 370–378.

Noroozi, M., Favaro, P., 2016. Unsupervised learning of visual representations by solving jigsaw puzzles. In: European Conference on Computer Vision. Springer, pp. 69–84.

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. In: European Conference on Computer Vision. Springer, pp. 762–780.

Ouyang, C., Biffi, C., Chen, C., Kart, T., Qiu, H., Rueckert, D., 2022. Self-supervised learning for few-shot medical image segmentation. IEEE Trans. Med. Imaging.

Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A., 2016. Context encoders: Feature learning by inpainting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2536–2544.

Peng, J., Pedersoli, M., Desrosiers, C., 2020. Mutual information deep regularization for semi-supervised segmentation. In: Medical Imaging with Deep Learning. PMLR, pp. 601–613.

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-supervised medical image segmentation with meta-labels. Adv. Neural Inf. Process. Syst. 34, 16686–16699.

Perez, F., Vasconcelos, C., Avila, S., Valle, E., 2018. Data augmentation for skin lesion analysis. In: OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis. Springer, pp. 303–311.

Pinheiro, P.O., Collobert, R., Dollár, P., 2015. Learning to segment object candidates. arXiv preprint arXiv:1506.06204.

Qin, Y., Zheng, H., Huang, X., Yang, J., Zhu, Y.M., 2019. Pulmonary nodule segmentation with CT sample synthesis using adversarial networks. Med. Phys. 46 (3), 1218–1229.

Rasmus, A., Berglund, M., Honkala, M., Valpola, H., Raiko, T., 2015. Semi-supervised learning with ladder networks. In: Advances in Neural Information Processing Systems. pp. 3546–3554.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 234–241.

Sajjadi, M., Javanmardi, M., Tasdizen, T., 2016. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In: Advances in Neural Information Processing Systems. pp. 1163–1171.

Seibold, C., Reiß, S., Kleesiek, J., Stiefelhagen, R., 2021. Reference-guided pseudo-label generation for medical semantic segmentation. arXiv preprint arXiv:2112.00735.

Shin, H.C., Tenenholtz, N.A., Rogers, J.K., Schwarz, C.G., Senjem, M.L., Gunter, J.L., Andriole, K.P., Michalski, M., 2018. Medical image synthesis for data augmentation and anonymization using generative adversarial networks. In: International Workshop on Simulation and Synthesis in Medical Imaging. Springer, pp. 1–11.

Sohn, K., Berthelot, D., Li, C.L., Zhang, Z., Carlini, N., Cubuk, E.D., Kurakin, A., Zhang, H., Raffel, C., 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. arXiv preprint arXiv:2001.07685.

Tarvainen, A., Valpola, H., 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. arXiv preprint arXiv:1703.01780.

Tschannen, M.T., Djolonga, J., Ritter, M., Mahendran, A., Houlsby, N., Gelly, S., Lučić, M., 2020. Self-supervised learning of video-induced visual invariances. In: Conference on Computer Vision and Pattern Recognition. URL: https://arxiv.org/abs/1912.02783.

Tustison, N.J., Avants, B.B., Cook, P.A., Zheng, Y., Egan, A., Yushkevich, P.A., Gee, J.C., 2010. N4ITK: improved N3 bias correction. IEEE Trans. Med. Imaging 29 (6), 1310–1320.

Valvano, G., Leo, A., Tsaftaris, S.A., 2021. Learning to segment from scribbles using multi-scale adversarial attention gates. IEEE Trans. Med. Imaging.

Van Gansbeke, W., Vandenhende, S., Georgoulis, S., Van Gool, L., 2021. Unsupervised semantic segmentation by contrasting object mask proposals. arXiv preprint arXiv:2102.06191.

Verma, V., Lamb, A., Kannala, J., Bengio, Y., Lopez-Paz, D., 2019. Interpolation consistency training for semi-supervised learning. In: International Joint Conference on Artificial Intelligence. pp. 3635–3641.

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

Wang, X., Zhang, R., Shen, C., Kong, T., Li, L., 2020. Dense contrastive learning for self-supervised visual pre-training. arXiv preprint arXiv:2011.09157.

Wang, W., Zhou, T., Yu, F., Dai, J., Konukoglu, E., Van Gool, L., 2021. Exploring cross-image pixel contrast for semantic segmentation. arXiv preprint arXiv:2101.11939.

Wu, Z., Xiong, Y., Yu, S.X., Lin, D., 2018. Unsupervised feature learning via non-parametric instance discrimination. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3733–3742.

Wu, Y., Xu, M., Ge, Z., Cai, J., Zhang, L., 2021. Semi-supervised left atrium segmentation with mutual consistency training. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 297–306.

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, Q., Luong, M.T., Hovy, E., Le, Q.V., 2020a. Self-training with noisy student improves imagenet classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10687–10698.

Xie, Y., Zhang, J., Liao, Z., Xia, Y., Shen, C., 2020b. PGL: Prior-guided local self-supervised learning for 3D medical image segmentation. arXiv preprint arXiv:2011.12640.

You, C., Yang, J., Chapiro, J., Duncan, J.S., 2020. Unsupervised wasserstein distance guided domain adaptation for 3D multi-domain liver segmentation. In: Interpretable and Annotation-Efficient Learning for Medical Image Computing. Springer, pp. 155–163.

You, C., Zhao, R., Liu, F., Chinchali, S., Topcu, U., Staib, L., Duncan, J.S., 2022. Class-aware generative adversarial transformers for medical image segmentation. arXiv preprint arXiv:2201.10737.

You, C., Zhao, R., Staib, L., Duncan, J.S., 2021a. Momentum contrastive voxel-wise representation learning for semi-supervised volumetric medical image segmentation. arXiv preprint arXiv:2105.07059.

You, C., Zhou, Y., Zhao, R., Staib, L., Duncan, J.S., 2021b. Simcvd: Simple contrastive voxel-wise representation distillation for semi-supervised medical image segmentation. arXiv preprint arXiv:2108.06227.

Yu, L., Wang, S., Li, X., Fu, C.W., Heng, P.A., 2019a. Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 605–613.

Yu, B., Zhou, L., Wang, L., Shi, Y., Fripp, J., Bourgeat, P., 2019b. Ea-GANs: edge-aware generative adversarial networks for cross-modality MR image synthesis. IEEE Trans. Med. Imaging 38 (7), 1750–1762.

Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D., 2017a. mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412.

Zhang, R., Isola, P., Efros, A.A., 2016. Colorful image colorization. In: European Conference on Computer Vision. Springer, pp. 649–666.

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

Zhang, Z., Yang, L., Zheng, Y., 2018. Translating and segmenting multimodal medical volumes with cycle-and shape-consistency generative adversarial network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 9242–9251.

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

Zhao, X., Vemulapalli, R., Mansfield, P., Gong, B., Green, B., Shapira, L., Wu, Y., 2020. Contrastive learning for label-efficient semantic segmentation. arXiv preprint arXiv:2012.06985.

Zheng, H., Lin, L., Hu, H., Zhang, Q., Chen, Q., Iwamoto, Y., Han, X., Chen, Y.W., Tong, R., Wu, J., 2019. Semi-supervised segmentation of liver using adversarial learning with deep atlas prior. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 148–156.

Zhou, Y., Chen, H., Lin, H., Heng, P.A., 2020. Deep semi-supervised knowledge distillation for overlapping cervical cell instance segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 521–531.

Zhou, Y., Xu, H., Zhang, W., Gao, B., Heng, P.A., 2021. C3-SemiSeg: Contrastive semi-supervised segmentation via cross-set learning and dynamic class-balancing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7036–7045.