Pseudo-label Guided Contrastive Learning for Semi-supervised Medical Image Segmentation

Hritam Basak
Stony Brook University, NY, USA
hbasak@cs.stonybrook.edu

Zhaozheng Yin
Stony Brook University, NY, USA
zyin@cs.stonybrook.edu

Abstract

Although recent works in semi-supervised learning (SemiSL) have accomplished significant success in natural image segmentation, the task of learning discriminative representations from limited annotations has been an open problem in medical images. Contrastive Learning (CL) frameworks use the notion of similarity measure which is useful for classification problems, however, they fail to transfer these quality representations for accurate pixel-level segmentation. To this end, we propose a novel semi-supervised patch-based CL framework for medical image segmentation without using any explicit pretext task. We harness the power of both CL and SemiSL, where the pseudo-labels generated from SemiSL aid CL by providing additional guidance, whereas discriminative class information learned in CL leads to accurate multi-class segmentation. Additionally, we formulate a novel loss that synergistically encourages inter-class separability and intra-class compactness among the learned representations. A new inter-patch semantic disparity mapping using average patch entropy is employed for a guided sampling of $positives$ *and* $negatives$ *in the proposed CL framework. Experimental analysis on three publicly available datasets of multiple modalities reveals the superiority of our proposed method as compared to the state-of-the-art methods. Code is available at: GitHub.*

1. Introduction

Accurate segmentation of medical images provides salient and insightful information to clinicians for appropriate diagnosis, disease progression, and proper treatment planning. With the recent emergence of neural networks, supervised deep learning approaches have achieved state-of-the-art performance in multiple medical image segmentation tasks [11, 36, 41]. This can be attributed to the availability of large annotated datasets. But, obtaining pixel-wise annotations in a large scale is often time-consuming, requires expertise, and incurs a huge cost, thus methods alleviating these requirements are highly expedient.

Semi-supervised learning (SemiSL) based methods are promising directions to this end, requiring a very small amount of annotations, and producing pseudo-labels for a large portion of unlabeled data, which are further utilized to train the segmentation network [32, 33]. In recent years, these methods have been widely recognized for their superior performance in downstream tasks (like segmentation, object detection, etc.), not only in natural scene images but also in biomedical image analysis [3, 4, 64]. Traditional SemiSL methods employ regression, pixel-wise cross entropy (CE), or mean squared error (MSE) loss terms or their variants. But, none of these losses imposes intra-class compactness and inter-class separability, restricting their full learning potential. Recent SemiSL methods in medical vision employing self-ensembling strategy [14, 44] have received attention because of their state-of-the-art performance in segmentation tasks. However, they are designed for a single dataset, failing to generalize across domains.

Unsupervised domain adaptation (UDA) [18, 61] can be utilized to address this problem, e.g., Xie *et al*. [60] proposed an efficient UDA method with self-training strategy to unleash the learning potential. However, most of these methods heavily rely upon abundant source labels, hence producing substandard performance with limited labels in clinical deployment [71]. Representational learning is another promising way to learn from limited annotations, where models trained for pretext tasks on large source domains can be transferred for downstream tasks in the target domain. Current advancements in representational learning have been ascribed as the upturn of contrastive learning (CL) [23], that aims to distinguish similar samples ($positive$) from dissimilar ones ($negative$) regarding a specified anchor point in a projected embedding space. This idea has resulted in substantial advancements in self-supervision paradigms by learning useful representations from large-scale unlabeled data [9, 43, 57]. The fundamental idea of CL is to $pull$ the semantically similar samples together and $push$ the dissimilar ones apart in the embedding space. This is accomplished by suitably designing an objective function, also known as the Contrastive Loss function, which optimizes the mutual information amongst different data points. The learned information from the pretext task can thereafter be transferred for downstream tasks such as classification [62], segmentation [53, 66], etc.

Despite their great success in recent years, CL frameworks are not devoid of problems, which broadly include: **(a)** sampling bias and aggravated **class collision** are reported in [15] because semantically similar instances are forcefully contrasted due to unguided selection of $negative$ samples [9], causing substandard performance; **(b)** as suggested in [21], it is a common and desirable practice in CL to adapt a model trained for some pretext task on an existing large-scale dataset of source domain (e.g., ImageNet) to a specific downstream task of the target domain. However, significant **domain shifts** in heterogeneous datasets may often hurt the overall performance [73], especially in medical images; and **(c)** designing a suitable **pretext task** can be challenging, and often cannot be generalized across datasets [37]. The first of these problems can be addressed by having access to labeled samples. For instance, [27] shows that including labels significantly improves the classification performance, but this is in a fully supervised setting. There have been recent attempts to partially address the last two problems, which are highlighted in section 2.

Our Proposal and Contribution

Taking motivation from these unsolved problems, we aim to leverage the potential of CL in the realm of SemiSL through several novel contributions:

- We propose a novel end-to-end segmentation paradigm by harnessing the power of both CL and SemiSL. In our case, the pseudo-labels generated in SemiSL aids CL by providing an additional guidance to the metric learning strategy, whereas the important class discriminative feature learning in CL boosts the multi-class segmentation performance of SemiSL. Thus **SemiSL aids CL and vice-versa** in medical image segmentation tasks.
- We introduce a novel Pseudo-label Guided Contrastive Loss (**PLGCL**) which can mine class-discriminative features without any explicit **training on pretext tasks**, thereby demonstrating **generalizability across multiple domains**.
- We employ a patch-based CL framework, where the $positive$ and $negative$ patches are sampled from an entropy-based metric guided by the pseudo-labels obtained in the SemiSL setting. This prevents (**class collision**), i.e., forceful and unguided contrast of semantically similar instances in CL.
- Upon the evaluation on three datasets from different domains, our method is proven to be effective, adding to its **generalizability** and **robustness**.

2. Related Works

2.1. Semi-supervised Learning

SemiSL-based approaches extract useful representations from a large set of unlabeled samples in tandem with supervised learning on a few labeled samples. Strategies employed by existing SemiSL methods include pseudo-labeling [35, 42], consistency regularization [4, 26], entropy minimization [22, 49], etc. Pseudo-labeling-based methods employ model training on labeled data, followed by the generation of pseudo-labels on an unlabeled dataset. The quality of the generated pseudo-labels is then fine-tuned using uncertainty-guided refinement [50], random propagation [16], etc. As procuring pixel-wise annotations for semantic segmentation is costly, consistency-based approaches enforce consistent predictions for augmented input images [17] or augmented feature embeddings [39] without using annotations. Entropy minimization enforces the model to output low-entropy predictions on unlabeled data [20]. Holistic approaches also employ a combination of these methods for various tasks [6, 46].

Another widely used method in semi-supervised medical image segmentation is Mean Teacher [47], which encourages consistent predictions between the student and teacher models. It has been extended to multiple SemiSL algorithms in recent years. Yu *et al*. [65] proposes an uncertainty-guided mean teacher framework (UA-MT), combined with transformation consistency for improved performance. Wang *et al*. [52] proposes a triple-uncertainty guided mean teacher framework by defining two auxiliary tasks: reconstruction and prediction of signed distance field on top of the mean teacher network to aid the model learning distinctive features for better predictions. Hang *et al*. [22] employs a global-local structure-aware entropy minimization method on top of the mean teacher network. Self-training approaches [60, 66] incorporate additional information from predictions on unlabeled data that can be used to improve the model performance. However, most of the existing semi-supervised segmentation methods do not explicitly stress the inter-class separability issue and thus inadvertently limit their performance, which we seek to address in our proposed work.

2.2. Contrastive Learning

Recent years have witnessed several powerful (dis)similarity learning approaches that employ contrastive loss for various computer vision tasks [12, 13, 37, 40]. Most of the previous CL methods in segmentation are employed in self-supervised pre-training to design a powerful feature extractor, which is then transferred for downstream tasks [9, 54]. For generating $positive$ pairs, these approaches rely heavily on data augmentations as supported by [2, 67], although it is noteworthy that a large number of $negatives$ is crucial for the success of these methods [8]. Zhao *et al*. [69] devises a CL strategy to mine relational characteristics between image-level and patch-level representations. Recently, the advantages of cross-image contrastive learning for medical image segmentation are demonstrated by Wang *et al*. [55]. However, a major drawback of CL in such a scenario is the **class collision** problem [1, 72] – where semantically similar patches get forcefully contrasted due to the uninformed $negative$ selection of the naive CL objective. This considerably hurts segmentation performance in a multi-class scenario, as shown by [28]. Our work aims to alleviate this issue by proposing a novel integration of CL with consistency regularization in semi-supervised segmentation. Unlike Boserup *et al*. [7], which requires an additional confidence network, we utilize the pseudo-labels for an entropy-based sampling of $positive$ and $negative$ queries for contrastive learning.

Some of the recent advancements employ contrastive learning in semi-supervised settings [21, 25, 68], where a model trained on a pretext classification task can be effectively transferred for a segmentation task. However, none of them effectively utilizes the pseudo-labels from SemiSL for refining CL, and vice-versa. Moreover, the success of these methods relies upon the careful design of **pretext task** and **minimal domain shift** between the pretext task domain and final segmentation domain. We try to address these problems in this work by designing an end-to-end segmentation framework through an effective utilization of CL in SemiSL setting. Chaitanya *et al*. [10] proposes a local contrastive learning-based self-training strategy, directed by the pseudo-labels, which is closest to our work. However, it is unclear how their proposed pixel-level CL can learn discriminative features without careful selection of $positives$ and $negatives$. Besides, their method lacks any pseudo-label refinement strategy, which is fundamental for the quality of generated pseudo-labels and is directly correlated to the metric learning scheme. Moreover, their pixel-wise CL framework suffers from out-of-memory issues, limiting them to sub-sample a small portion of pixels and inhibiting the model to learn global information. To address most of these problems, we propose patch-wise contrastive learning, guided by the pseudo-labels, and jointly optimize the CL loss and consistency loss in SemiSL for learning feature representations and refining the pseudo-labels simultaneously.

3. Proposed Method

Given a labeled image set $\mathbb{I}_L$ with its corresponding label set $\mathbb{Y}_L$ and an unlabeled image set $\mathbb{I}_U$ which contain $\mathcal{N}_L$ and $\mathcal{N}_U$ numbers of images, respectively (where $\mathcal{N}_L << \mathcal{N}_U$), we introduce a patch-wise contrastive learning strategy, guided by pseudo-labels, which aims to learn information from both $\mathbb{I}_L$ and $\mathbb{I}_U$. Our proposed method can be described in four steps: first, we define the generation of patches, directed by the effective utilization of (true or pseudo) labels (subsection 3.1), then we formulate a new contrastive loss function (subsection 3.2). After that, we define the overall learning objective (subsection 3.3), and finally, we describe the pseudo-label generation and refinement strategy in subsection 3.4.

3.1. Class-aware Patch Sampling

Let's represent the i^{th} image of a mini-batch as I_i, containing M pixels, where the m^{th} pixel in the image is denoted by $I_i(m); m \in [1, M]$. Our proposed framework uses an encoder and decoder network $\mathcal{E}_S$ and $\mathcal{D}_S$, parameterized by $\theta_{\mathcal{E},S}$ and $\theta_{\mathcal{D},S}$ respectively, to generate pseudo-label Y_i' from I_i[1], which is equivalently represented as class-confidence metric C_i, i.e., $[\mathcal{E}_S, \mathcal{D}_S] : I_i \rightarrow C_i$. Here $C_i = \{C_i^k(m)\}$ and $C_i^k(m)$ denotes the confidence of pixel m of image I_i belonging to class k, where $k = \{1, 2, ..K\}$ and $K(\geq 1)$ indicates the number of classes in a segmentation map. This confidence map is thereafter multiplied with I_i to obtain the $attended$ image $I_i'^k = I_i \odot C_i^k$, where $(\odot)$ indicates element-wise multiplication. This attended image is subjected to the generation of patches, where the j^{th} patch of the i^{th} attended image for the k^{th} class is denoted by $P_{i,j}^k$.

Given an anchor patch from class k, all the patches containing an object (or some part of it) of class k, are treated as $positive$, and all the patches from other $(K-1)$ classes are $negative$. Appropriate sampling of numerous patches is of utmost importance for CL. We can sample patches based on their class confidences, e.g., the average confidence of a patch $P_{i,j}^k$ is computed as:

$$Avg_{i,j}^k = \frac{\sum\limits_{m \in P_{i,j}^k} C_i^k(m)}{|P_{i,j}^k|}. \tag{1}$$

A high average patch confidence indicates patch $P_{i,j}^k$ is more likely to contain the object (or part of it) belonging to class k, whereas values close to 0 indicate the opposite. Values in-between indicate uncertainty in either direction. However, $Avg_{i,j}^k$ is just based on a patch's confidence on class k and it ignores two important items: (i) the patch's intensity appearance information and (ii) the confidence uncertainty between class k and other $(K-1)$ classes. Therefore, we propose to compute the average patch entropy based on the attended image $I_i'^k$. For a patch $P_{i,j}^k$, its average patch entropy is calculated from the pixels' intensity values in the attended image $I_i'^k$, expressed as:

[1] In case of a labeled sample $I_i \in \mathbb{I}_L$, the available ground truth $(Y_i \in \mathbb{Y}_L)$ is used instead of generating its pseudo-label Y_i'.

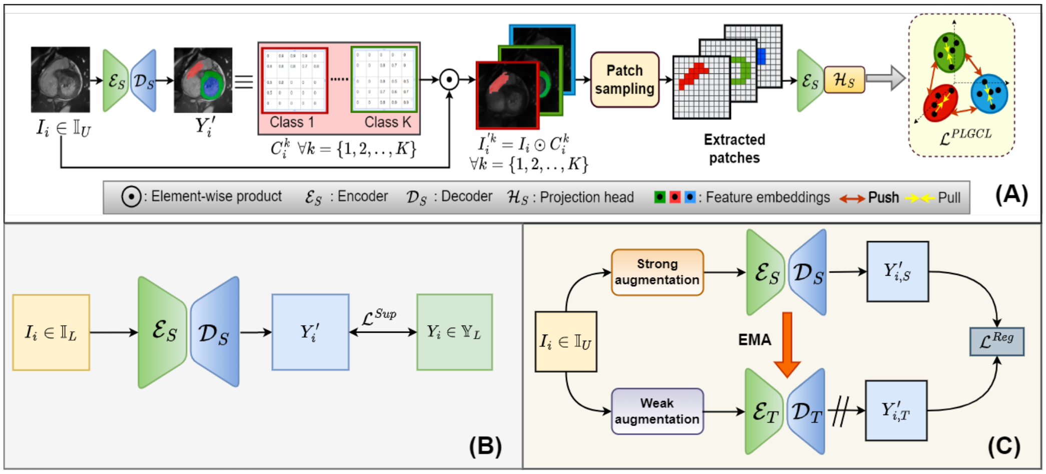

Figure 1. Overall workflow of our proposed method. (A) Our proposed pseudo-label guided contrastive learning strategy (details can be found in subsection 3.1 and subsection 3.2); (B) Student encoder $\mathcal{E}_S$ and decoder $\mathcal{D}_S$ are learned from labeled images $I_i \in \mathbb{I}_L$ using supervised loss $\mathcal{L}^{Sup}$; (C) Regularization loss $\mathcal{L}^{Reg}$ is computed between the prediction of the teacher model (with encoder $\mathcal{E}_T$ and decoder $\mathcal{D}_T$) on a weakly augmented version of unlabeled images $I_i \in \mathbb{I}_U$ and the prediction of student model on strongly augmented version. The weights of $\mathcal{E}_T$ and $\mathcal{D}_T$ are updated using the exponential moving average (EMA) from the student branch.

$$Ent_{i,j}^k = \frac{\sum_{m \in P_{i,j}^k} \mathcal{F}(I_i'^k(m))}{|P_{i,j}^k|}, \text{ where} \tag{2}$$

$$\mathcal{F}(x) = -x \log(x) - (1-x) \log(1-x) \tag{3}$$

is the entropy function, $I_i'^k(m)$ is the intensity value of the m^{th} pixel inside the patch. $Ent_{i,j}^k$ reflects three types of information of a patch j in image i: confidence belonging to class k, uncertainty regarding the other $(K-1)$ classes, and the intensity appearance from I_i (Note, $I_i'^k = I_i \odot C_i^k$). Thus, for a given anchor patch of class k, all the patches with n-nearest Ent_i^k values are considered as $positive$, and the rest as $negative$. These patches are passed through encoder $\mathcal{E}_S$ and projection head $\mathcal{H}_S$ to obtain the feature embeddings, which are then used for contrastive loss formulation in the following section. The embedding of an anchor point is considered as $query$, which is contrasted with all the other embeddings from other patches (considered as $keys$), which is the basis of our CL. This overall pipeline is shown in Figure 1(A).

3.2. Pseudo-label Guided Contrastive Loss

We propose a novel Pseudo-label Guided Contrastive Loss (PLGCL), assuming the availability of pseudo-label $\mathbb{Y}'_U$ for the unlabeled set $\mathbb{I}_U$ (the pseudo-label generation will be explained in subsection 3.4) along with the labeled samples $(\mathbb{I}_L, \mathbb{Y}_L)$. Previous works such as JCL [8] compute the expectation of the InfoNCE loss [38] over a distribution of $positive$ samples only, for a given query. In our case, due to the presence of class information in terms of class-wise patches, one can take the expectation of InfoNCE over the joint distribution of the class conditionals of both the $positive$ and $negative$ keys, which is the basis of PLGCL.

Let the u^{th} query patch of class k be denoted as P_u, and its corresponding v^{th} key patch is P_v^{k+} if v is $positive$ (i.e., it has the same pseudo/true class as patch P_u); otherwise, it is denoted as P_v^{k-} ($negative$ key patch of a class different from k). We denote the embeddings of P_u, P_v^{k+}, P_v^{k-} as f_u, f_v^{k+}, f_v^{k-} respectively, such that $\{f_u, f_v^{k+}, f_v^{k-}\} \leftarrow \mathcal{H}_S(\mathcal{E}_S(\{P_u, P_v^{k+}, P_v^{k-}\}))$. Let $f_v^{k+} \sim p(\cdot|k_+)$ and $f_v^{k-} \sim p(\cdot|k_-)$, the expectation of the InfoNCE loss with respect to the joint distribution $\mathcal{J}$, over all the class conditional densities $p(\cdot|k_+)$ and $p(\cdot|k_-)$, is expressed as:

$$L = -\mathbf{E}_{\mathcal{J}} \log \frac{\exp(f_u^T \cdot f_v^{k+}/\tau)}{\exp(f_u^T \cdot f_v^{k+}/\tau) + \sum_{k_-}\sum_{v} \exp(f_u^T \cdot f_v^{k-}/\tau)} \tag{4}$$

where τ is the temperature parameter [12]. Closed-form upper-bound of Equation 4 can be derived as:

$$\begin{aligned} L = \mathbf{E}_{\mathcal{J}} \Bigg[\log \Bigg(& \exp(f_u^T \cdot f_v^{k+}/\tau) \\ & + \sum_{k_-}\sum_{v} \exp(f_u^T \cdot f_v^{k-}/\tau) \Bigg) \Bigg] - \mathbf{E}_{p(\cdot|k_+)} \left(f_u^T \cdot f_v^{k+}/\tau \right) \\ \leq \log & \left[\mathbf{E}_{\mathcal{J}} \left(\exp(f_u^T \cdot f_v^{k+}/\tau) + \sum_{k_-}\sum_{v} \exp(f_u^T \cdot f_v^{k-}/\tau) \right) \right] \\ & - f_u^T \mathbf{E}_{p(\cdot|k_+)} \left(f_v^{k+}/\tau \right) \end{aligned}$$

The last equation is obtained using Jensen inequality on concave function, i.e., $\mathbf{E}[\log(\cdot)] \leq \log[\mathbf{E}(\cdot)]$. Now, using Gaussianity assumption [8] over all the class conditional densities $p(\cdot|k_+)$ and $p(\cdot|k_-)$, we parameterize them as $f_v^{k+} \sim Norm(\mu_{f_v^{k+}}, \sigma_{f_v^{k+}})$ and $f_v^{k-} \sim Norm(\mu_{f_v^{k-}}, \sigma_{f_v^{k-}})$, where μ and σ represent the mean and covariance matrix, respectively. Leveraging $\mathbf{E}_x(e^{a^T x}) = e^{a^T\mu + \frac{1}{2}a^T\sigma a}$ when $x \sim Norm(\mu, \sigma)$, and $\mathbf{E}_{g(a,b,c,..)}h(a) = \mathbf{E}_{g(a)}h(a)$, the upper bound of Equation 4 leads to our patch-wise pseudo-label guided contrastive loss:

$$\mathcal{L}_u^{PLGCL} = \log\Bigg[\exp\left(f_u^T\mu_{f_v^{k+}}/\tau + \frac{\lambda}{2\tau^2}f_u^T\sigma_{f_v^{k+}}f_u\right) + \zeta\sum_{k_-} exp\left(f_u^T\mu_{f_v^{k-}}/\tau + \frac{\lambda}{2\tau^2}f_u^T\sigma_{f_v^{k-}}f_u\right)\Bigg] - f_u^T\mu_{f_v^{k+}}/\tau \tag{5}$$

where ζ is a scaling factor that originates from the term $\sum_v$, i.e., summation over all the *negative* embeddings for a particular class. As stated in [8], the statistics are more informative in the later stage of training, hence λ is used to scale the effect of $\sigma_{f^{k_+}}$ that stabilizes the training. The proposed loss $\mathcal{L}^{PLGCL}$ relies upon reasonable estimation of $\mu_{f_v^{k+}}, \sigma_{f_v^{k+}}, \mu_{f_v^{k-}}, \sigma_{f_v^{k-}}$ from f_v^{k+}, f_v^{k-}. We address this problem by accurate estimation of *positives* and *negatives* based on an entropy-based sampling strategy (subsection 3.1).

3.3. The Overall Learning Objective

Along with the proposed CL framework, our method can mine important pixel-level information from the images in a semi-supervised setting, for which we employ a student-teacher network [47]. We represent the student encoder and decoder as $\mathcal{E}_S, \mathcal{D}_S$, parameterized by $\theta_{\mathcal{E},S}, \theta_{\mathcal{D},S}$, respectively, and the teacher encoder-decoder model $\mathcal{E}_T, \mathcal{D}_T$, parameterized by $\theta_{\mathcal{E},T}, \theta_{\mathcal{D},T}$. Let the student projection head be denoted as $\mathcal{H}_S$, parameterized by $\theta_{\mathcal{H},S}$. With the student-teacher network, we define the consistency cost for an unlabeled image $I_i \in \mathbb{I}_U$ as the cross entropy (CE) loss between the outputs of student and teacher models as:

$$\mathcal{L}_i^{Reg} = CE\Big[\mathcal{D}_S\Big(\mathcal{E}_S(I_i^s)\Big), \mathcal{D}_T\Big(\mathcal{E}_T(I_i^w)\Big)\Big] \tag{6}$$

where I_i^s and I_i^w represent the strong and weak augmentations of input I_i. Additionally, we compute the supervised CE loss between the prediction of labeled samples $I_i \in \mathbb{I}_L$ from the student encoder-decoder network and the available ground truths $Y_i \in \mathbb{Y}_L$ as:

$$\mathcal{L}_i^{Sup} = CE\Big[\mathcal{D}_S\Big(\mathcal{E}_S(I_i)\Big), Y_i\Big] \tag{7}$$

The final objective function is boiled down to:

$$\mathcal{L}_i^{total} = \frac{1}{|\mathcal{B}_L|}\sum_{I_i\in\mathcal{B}_L}\mathcal{L}_i^{Sup} + \beta\frac{1}{|\mathcal{B}_U|}\sum_{I_i\in\mathcal{B}_U}\mathcal{L}_i^{Reg} + \gamma\frac{1}{|\mathcal{B}|}\sum_{I_i\in\mathcal{B}}\mathcal{L}_i^{PLGCL} \tag{8}$$

where $\mathcal{B}$ is the sampled mini-batch; $\mathcal{B}_L, \mathcal{B}_U$ are the labeled and unlabeled samples in the mini-batch, respectively, and $|\cdot|$ is the set cardinality. During training, the student network parameters are updated by minimizing Equation 8 using the SGD optimizer whereas the teacher network parameters are updated using exponential moving average (EMA) as:

$$\theta_{\mathcal{E},T}(t+1) \leftarrow \alpha\theta_{\mathcal{E},T}(t) + (1-\alpha)\theta_{\mathcal{E},S}(t+1) \tag{9}$$

$$\theta_{\mathcal{D},T}(t+1) \leftarrow \alpha\theta_{\mathcal{D},T}(t) + (1-\alpha)\theta_{\mathcal{E},S}(t+1) \tag{10}$$

where t tracks the step number, and α is the "smoothing coefficient" [47] or the "momentum coefficient" [23].

3.4. Pseudo-label Generation and Refinement

As shown in Figure 1, our method consists of three parts (A) pseudo-label guided contrastive learning, (B) consistency regularization for unlabeled samples, and (C) supervised learning for labeled samples. The contrastive learning part needs pseudo-labels as the input. To this end, we use a small semi-supervised warm-up phase for 50 epochs to generate the pseudo-labels using only $\mathcal{L}^{Reg}$ and $\mathcal{L}^{Sup}$ in Equation 8. A weak and strong augmentation of an image $I_i \in \mathbb{I}_U$ is generated and passed through the student and teacher models, respectively. We enforce the consistency between the two obtained outputs using the consistency loss $\mathcal{L}^{Reg}$ (refer Equation 6). Additionally, we also compute the supervised CE loss $\mathcal{L}^{Sup}$ between the segmentation output of the student model Y_i' for image $I_i \in \mathbb{I}_L$ and the available ground truth $Y_i \in \mathbb{Y}_i$.

The warm-up training generates initial pseudo-labels, and then the contrastive loss $\mathcal{L}^{PLGCL}$ is introduced after the warm-up phase, and the model is trained with pseudo-labels being refined until convergence. The parameters of the student model are updated iteratively using the current network parameters and the gradient of the computed loss, whereas the teacher network parameters are updated using EMA from the student model (Equation 9 and Equation 10). The overall workflow is summarized in algorithm 1.

4. Experiments and Results

We evaluate the proposed method on three widely used datasets with various medical imaging modalities: MRI, CT, and histopathology.

Algorithm 1: Workflow of our proposed method.

Input: $\mathbb{I}_L, \mathbb{Y}_L, \mathbb{I}_U$
Warm-up $\theta_{\mathcal{E},S}, \theta_{\mathcal{D},S}, \theta_{\mathcal{E},T}, \theta_{\mathcal{D},T}$ using $\mathcal{L}^{Sup}$, $\mathcal{L}^{Reg}$
while $iteration \leq max\_iteration$ **do**
  Sample batch $\mathcal{B}$ from $\mathbb{I}_L \cup \mathbb{I}_U$
  /* Compute supervised loss */
  for $I_i \in \mathcal{B}_L$ **do**
    $Y_i' \leftarrow \mathcal{D}_S(\mathcal{E}_S(I_i))$
    $\mathcal{L}^{Sup} \leftarrow \text{CE}(Y_i', Y_i)$
  end for
  /* Compute regularization loss */
  for $I_i \in \mathcal{B}_U$ **do**
    $I_i^s \leftarrow$StrongAugment(I_i); $I_i^w \leftarrow$WeakAugment(I_i)
    $Y_{i,S}' \leftarrow \mathcal{D}_S(\mathcal{E}_S(I_i^s))$; $Y_{i,K}' \leftarrow \mathcal{D}_T(\mathcal{E}_T(I_i^w))$
    $\mathcal{L}^{Reg} \leftarrow CE(Y_{i,S}', Y_{i,T}')$
  end for
  /* Patch sampling */
  $[\mathcal{E}_S, \mathcal{D}_S] : I \rightarrow C^k$; $\forall k = \{1, 2, ..K\}$
  $I'^k \leftarrow I \odot C^k; \forall k = \{1, 2, .., K\}$
  Generate patches from I'^k
  /* Compute PLGCL loss */
  for all P_u, P_v^{k+}, P_v^{k-} **do**
    $\{f_u, f_v^{k+}, f_v^{k-}\} \leftarrow \mathcal{H}_S(\mathcal{E}_S(\{P_u, P_v^{k+}, P_v^{k-}\}))$
    $\mathcal{L}^{PLGCL} \leftarrow PLGJCL\left[f_u, f_v^{k+}, f_v^{k-}\right]$ (Equation 5)
  end for
  $\mathcal{L}^{total} \longleftarrow$
  $\frac{1}{|\mathcal{B}_L|} \sum\limits_{I_r \in \mathcal{B}_L} \mathcal{L}^{Sup} + \beta \frac{1}{|\mathcal{B}_U|} \sum\limits_{I_s \in \mathcal{B}_U} \mathcal{L}^{Reg} + \gamma \frac{1}{|\mathcal{B}|} \sum\limits_{I \in \mathcal{B}} \mathcal{L}^{PLGCL}$
  /* Update network parameters */
  $\theta_{\mathcal{E},S} \leftarrow \theta_{\mathcal{E},S} - \nabla_{\theta_{\mathcal{E},S}} \mathcal{L}_{total}$;$\theta_{\mathcal{D},S} \leftarrow \theta_{\mathcal{D},S} - \nabla_{\theta_{\mathcal{D},S}} \mathcal{L}_{total}$
  $\theta_{\mathcal{E},T} \leftarrow \alpha\theta_{\mathcal{E},T} + (1-\alpha)\theta_{\mathcal{E},S}$;$\theta_{\mathcal{D},T} \leftarrow \alpha\theta_{\mathcal{D},T} + (1-\alpha)\theta_{\mathcal{D},S}$
end while
Return $\theta_{\mathcal{E},S}, \theta_{\mathcal{D},S}, \theta_{\mathcal{E},T}, \theta_{\mathcal{D},T}$

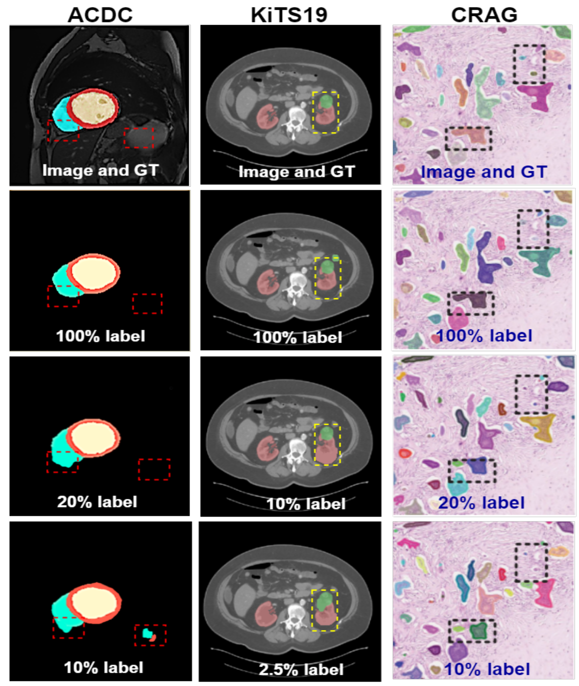

Figure 2. Visual comparison of segmentation results with different percentages of labeled data for training, 100% indicates fully-supervised setting.

4.1. Dataset

(1) ACDC dataset is a cardiac MRI dataset [5] that contains 100 short axis cine-MRIs, captured using 3T and 1.5T machines, and contains expert annotations for three classes: left and right ventricle (LV, RV), and myocardium (MYO). We followed the works [31, 57] to split the dataset into $70-10-20$ as the training, validation, and test sets, respectively. **(2) KiTS19** is a tumor segmentation dataset [24], containing 210 labeled volumes of kidney CT. We followed the experimental settings of [26], i.e., 150 for training, 20 for validation, and 40 for testing. **(3) Colorectal Adenocarcinoma Gland (CRAG) dataset** [19] contains 213 H&E WS histopathological images taken with an OmnyxVL120 scanner. It has images with 20x objective magnification with a resolution of 0.55 µm/pixel. We follow [43] to split the data into $80-10-10$ training, test, and validation ratio.

4.2. Implementation

Our proposed method is implemented in a PyTorch environment and executed using a Tesla V100 GPU with 32GB RAM. We use three different metrics for the evaluation of model performance, namely Dice Similarity Score (DSC), Hausdorff Distance 95 (HD95) and Average Symmetric Distance (ASD) [9]. For a fair comparison, we follow the previous SemiSL works [4, 10, 26] and use 10% and 20% labeled data for training the model, and the rest as unlabeled, except for KiTS19, where we follow the same training protocol as [51] to use 2.5% and 10% images as labeled while training. We use a simple U-Net [41] backbone for the encoder-decoder structure, and the projection head is basically a shallow FC layer [12]. The model is converged using an ADAM optimizer with a batch size of 16 and a learning rate of $1e-4$. τ and λ in Equation 5 are taken as 0.2 and 4, following [8]. α in Equation 9, β, γ in Equation 5, and n in the n-nearest entropy-based sampling in subsection 3.1 are set to 0.999, 0.25, 0.2, and 20, respectively by validation. For weak augmentations, we use random rotation and crop, and morphological and brightness changes are used for strong augmentation [63].

4.3. Results and Comparison with SoTA

We experiment with different percentages of labeled data and compare the performance with its counterpart trained

Method	labeled data (%)	Evaluation Metrics		
		DSC ↑	HD95 ↓	ASD ↓
UA-MT [65]	10%	0.816	12.35	3.62
Double-UA [56]		0.833	5.31	1.92
MC-Net [59]		0.863	7.08	2.08
MC-Net+ [58]		0.871	6.68	2.00
SASSNet [29]		0.841	5.03	1.40
DTC [32]		0.827	10.81	2.99
LCLPL [10]		0.881	5.11	1.81
Ours		0.891	4.98	1.80
UA-MT [65]	20%	0.857	4.06	1.54
URPC [34]		0.851	4.26	1.77
MC-Net [59]		0.878	3.91	1.52
MC-Net+ [58]		0.885	4.35	1.54
SASSNet [29]		0.871	5.84	2.15
DTC [32]		0.863	6.14	2.11
LCLPL [10]		0.905	3.91	1.51
Ours		0.912	3.82	1.49
Supervised	100%	0.923	3.66	1.41

(a) **ACDC**

Method	labeled data (%)	Evaluation Metrics		
		DSC ↑	HD95 ↓	ASD ↓
UA-MT [65]	2.50%	0.871	11.74	3.56
SASSNet [29]		0.888	8.32	2.34
CoraNet [45]		0.882	8.21	2.44
DTC [32]		0.885	7.99	2.40
GBDL [51]		0.898	6.85	1.78
Triple-UA [52]		0.878	7.94	2.42
Double-UA [56]		0.887	8.04	2.34
Ours		0.905	6.75	1.75
UA-MT [65]	10%	0.883	9.46	2.89
SASSNet [29]		0.891	7.54	2.51
CoraNet [45]		0.898	7.23	1.81
DTC [32]		0.894	7.31	1.91
GBDL [51]		0.911	6.38	1.51
Triple-UA [52]		0.887	7.55	2.12
Double-UA [56]		0.895	7.42	2.16
Ours		0.919	6.32	1.51
Supervised	100%	0.934	6.10	1.44

(b) **KiTS19**

Method	labeled data (%)	Evaluation Metrics		
		DSC↑	HD95↓	ASD↓
ICT [48]	10%	0.862	1.52	2.39
Double-UA [56]		0.877	1.45	2.56
HCE [26]		0.874	1.31	2.44
DTC [32]		0.841	1.81	2.61
TCSM [30]		0.853	1.52	2.46
UA-MT [65]		0.816	1.89	2.58
Ours		0.882	1.50	2.42
ICT [48]	20%	0.866	1.46	2.22
Double-UA [56]		0.883	1.28	2.06
HCE [26]		0.885	1.23	2.11
DTC [32]		0.859	1.70	2.24
TCSM [30]		0.877	1.41	2.36
UA-MT [65]		0.856	1.69	2.13
Ours		0.891	1.24	2.01
Supervised	100%	0.911	1.19	1.88

(c) **CRAG**

Table 1. Comparison of our method with state-of-the-art semi-supervised segmentation methods on three datasets. Values highlighted in RED and GREEN indicate the best and second best results among all the SemiSL methods compared. Note that the evaluation is at pixel-level and the three datasets have 10^7, 10^7, 10^8 pixels in their testing sets, respectively.

in a fully-supervised manner (i.e., 100% of labels used) in Table 1. Qualitative analysis of the results using different label percentages is depicted in Figure 2. As observed in the last two rows of Table 1 and Figure 2, our method can mine discriminative features by using very few labels, leading to good very close results to the fully-supervised counterpart.

Next, our proposed method is compared with the existing state-of-the-art CL and SemiSL-based segmentation methods. As shown in Table 1(a), our proposed method outperforms all the SoTA SemiSL methods like UA-MT [65], URPC [34], DTC [32], MC-Net [59], SASSNet [29] on the MRI dataset. As discussed in section 2, LCLPL [10] proposes a pseudo-label guided local contrastive learning, which is closest to our work. However, their method suffers from the unguided selection of $positives$ and $negatives$, without pseudo-label refinement, leading to sub-optimal performance. In contrast, our method benefits from the proposed PLGCL loss and entropy-based patch sampling, resulting in enhanced performance. Moreover, these margins are larger with fewer labels (10%), indicating the robustness of our method to learn from limited annotations. Similar observations are made for the KiTS19 dataset, where it is evident from Table 1(b) that the proposed method outperforms the widely used SemiSL methods like [29,45,56,65]. One of the recent methods [51], produces the second best result using a generative Bayesian deep learning strategy in SemiSL, lacking the capability to mine class information and address class-collision. Most of the other methods lack any feedback mechanism for the teacher network by observing how pseudo-labels would affect the student. In our case, however, the regularization network benefits from the CL framework, and vice versa, resulting in the best performance, even by using only 2.5% labels. In Table 1(c), we compare the performance of our work with the existing SoTA methods on CRAG dataset. In this case, some recent methods like Double-UA [56], DTC [32], UA-MT [65] produce good results, but fail to generalize in different modalities, making our method a clear winner in all three datasets.

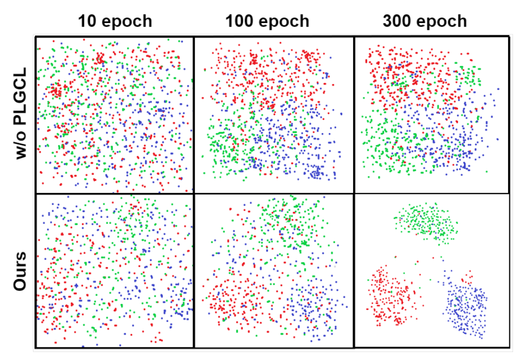

Figure 3. Ablation experiment: t-SNE decomposition of representation space produced by encoders $\mathcal{E}_S$ and projection head $\mathcal{H}_S$ at different training stages on **ACDC** dataset (20% labeled) with and without the proposed PLGCL.

4.4. Ablation Study

We perform a set of ablation experiments to validate the effectiveness of individual components.

4.4.1 Effectiveness of PLGCL

We perform experimentation with and without the pseudo-label guided contrastive loss ($\mathcal{L}^{PLGCL}$). As shown in Table 2, removing PLGCL affects the performance significantly as it helps the model learn discriminative class information, hence the introduction of PLGCL improves segmentation performance. Moreover, it is so powerful that

Method		ACDC			KiTS19			CRAG		
Warm-up	PLGCL	**DSC↑**	**HD95↓**	**ASD↓**	**DSC↑**	**HD95↓**	**ASD↓**	**DSC↑**	**HD95↓**	**ASD↓**
×	×	0.799	8.77	4.44	0.831	8.04	3.11	0.813	2.36	3.44
✓	×	0.822	7.54	3.61	0.855	7.72	2.62	0.819	2.04	3.52
×	✓	0.885	5.21	2.04	0.901	6.41	1.81	0.873	1.64	2.53
✓	✓	0.891	4.98	1.80	0.919	6.32	1.51	0.882	1.50	2.42

Table 2. Ablation study on three different datasets (using 10% labeled data) to identify the contribution of individual components. The last row, highlighted in RED indicates our results.

Similarity metric	Label = 10%			Label = 20%		
	DSC↑	**HD95↓**	**ASD↓**	**DSC↑**	**HD95↓**	**ASD↓**
Cosine similarity	0.820	9.118	6.016	0.832	7.611	4.445
Class Confidence	0.873	5.091	2.878	0.877	4.497	2.014
Entropy (**ours**)	0.891	4.980	1.802	0.912	3.823	1.491

Table 3. Comparison of different similarity measures for patch sampling. Experiments are performed on the **ACDC** dataset.

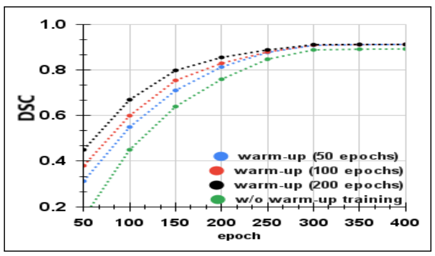

Figure 4. Visualization of warm-up phase ensuring speedy convergence and better overall performance on **ACDC** (20% labeled)

even without warm-up (i.e., start to use pseudo-labels immediately after the first training epoch), it can still help the model produce quite accurate segmentation performance (row 3 in Table 2). Besides, we also analyze the t-SNE decomposition of representation space with and without $\mathcal{L}^{PLGCL}$, as shown in Figure 3. It is interesting to observe how PLGCL results in improved clustering of feature embeddings as the training progresses, yielding good inter-class separability and intra-class compactness. On the other hand, without PLGCL, the embeddings from various classes entangle with each other in the feature space. This sufficiently demonstrates the effectiveness of our proposed scheme to address the critical issue of class collision in CL.

4.4.2 Effectiveness of Warm-up Training

In the pseudo-label generation and refinement, we use a small warm-up phase using only $\mathcal{L}^{Sup}$ and $\mathcal{L}^{Reg}$, followed by a full model training. To identify the effectiveness of warm-up, we perform two sets of experiments with and without warm-up. First, the model is warmed-up and the generated pseudo-labels after this are utilized for CL and are refined iteratively during the full model training. In the second experiment, we directly use the pseudo-labels from the first iteration for CL without any iterative refinement. As shown in Figure 4, warm-up helps the model initialize better for the second phase of training, which is also corroborated by [70]. Warm-up for longer period, although provides initial boost, does not necessarily improves the final segmentation performance (refer Figure 4). Better initialization provides a meaningful additional signal for strong guidance to PLGCL, which is evident from the observations in Table 2, where introducing warm-up along with PLGCL improves the performance by ($\sim 7-10\%$) throughout.

4.4.3 Effectiveness of Patch Sampling

We compare our patch sampling method with two noteworthy ones: **(A) Cosine similarity:** It is the most obvious and common metric for similarity measurement between two patches. Given two vectorized patches a and b, the cosine similarity is calculated as: $Sim(a,b) = a \cdot b/|a||b|$. **(B) Class Confidence:** For a patch $P^k_{i,j}$, we calculate the average patch confidence $Avg^k_{i,j}$ (Equation 1), and patches having similar confidence values are sampled as $positives$, and the rest as $negatives$. Although simple, the cosine-similarity-based patch-sampling from $I_i^{'k}$ fails to produce satisfactory results as shown in Table 3. Class confidence-based sampling, however, performs better. As the sampling sets of $positive$ and $negative$ are not always disjoint, they can lead to a higher misclassification rate, resulting in sub-optimal performance. We argue that it is better to sample the $positives$ and $negatives$ based on the entropy in the image attended by the class confidence map Equation 2 as it is a better metric for disparity mapping among patches.

5. Conclusion

In this work, we formulate a new CL strategy in a SemiSL setting by the effective utilization of pseudo-labels. To the best of our knowledge, this is the first attempt to integrate CL in a semi-supervised setting using consistency regularization and pseudo-labeling for semi-supervised medical image segmentation. The proposed modality-agnostic model, when evaluated on three medical segmentation datasets from multiple domains, outperforms the SoTA methods, justifying its effectiveness and generalizability.

References

[1] Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. A theoretical analysis of contrastive unsupervised representation learning. *arXiv preprint arXiv:1902.09229*, 2019. 3

[2] YM Asano, C Rupprecht, and A Vedaldi. A critical analysis of self-supervision, or what we can learn from a single image. In *International Conference on Learning Representations*, 2019. 3

[3] 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. 1

[4] Hritam Basak, Sagnik Ghosal, and Ram Sarkar. Addressing class imbalance in semi-supervised image segmentation: A study on cardiac mri. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 224–233. Springer, 2022. 1, 2, 6

[5] 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*, 37(11):2514–2525, 2018. 6

[6] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32, 2019. 2

[7] Nicklas Boserup and Raghavendra Selvan. Efficient self-supervision using patch-based contrastive learning for histopathology image segmentation. *arXiv preprint arXiv:2208.10779*, 2022. 3

[8] Qi Cai, Yu Wang, Yingwei Pan, Ting Yao, and Tao Mei. Joint contrastive learning with infinite possibilities. *Advances in Neural Information Processing Systems*, 33:12638–12648, 2020. 3, 4, 5, 6

[9] Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu. Contrastive learning of global and local features for medical image segmentation with limited annotations. *Advances in Neural Information Processing Systems*, 33:12546–12558, 2020. 1, 2, 6

[10] Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu. Local contrastive loss with pseudo-label based self-training for semi-supervised medical image segmentation. *arXiv preprint arXiv:2112.09645*, 2021. 3, 6, 7

[11] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 1

[12] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2, 4, 6

[13] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255, 2020. 2

[14] Veronika Cheplygina, Marleen de Bruijne, and Josien PW Pluim. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Medical image analysis*, 54:280–296, 2019. 1

[15] Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. *Advances in neural information processing systems*, 33:8765–8775, 2020. 2

[16] Deng-Ping Fan, Tao Zhou, Ge-Peng Ji, Yi Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Inf-net: Automatic covid-19 lung infection segmentation from ct images. *IEEE Transactions on Medical Imaging*, 39(8):2626–2637, 2020. 2

[17] Geoff French, Samuli Laine, Timo Aila, Michal Mackiewicz, and Graham Finlayson. Semi-supervised semantic segmentation needs strong, varied perturbations. *arXiv preprint arXiv:1906.01916*, 2019. 2

[18] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015. 1

[19] Simon Graham, Hao Chen, Jevgenij Gamper, Qi Dou, Pheng-Ann Heng, David Snead, Yee Wah Tsang, and Nasir Rajpoot. Mild-net: Minimal information loss dilated network for gland instance segmentation in colon histology images. *Medical image analysis*, 52:199–211, 2019. 6

[20] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. *Advances in neural information processing systems*, 17, 2004. 2

[21] Ran Gu, Jingyang Zhang, Guotai Wang, Wenhui Lei, Tao Song, Xiaofan Zhang, Kang Li, and Shaoting Zhang. Contrastive semi-supervised learning for domain adaptive segmentation across similar anatomical structures. *IEEE Transactions on Medical Imaging*, 2022. 2, 3

[22] Wenlong Hang, Wei Feng, Shuang Liang, Lequan Yu, Qiong Wang, Kup-Sze Choi, and Jing Qin. Local and global structure-aware entropy regularized mean teacher model for 3d left atrium segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 562–571. Springer, 2020. 2

[23] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 1, 5

[24] Nicholas Heller, Fabian Isensee, Klaus H Maier-Hein, Xiaoshuai Hou, Chunmei Xie, Fengyi Li, Yang Nan, Guangrui Mu, Zhiyong Lin, Miofei Han, et al. The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge. *Medical Image Analysis*, page 101821, 2020. 6

[25] Xinrong Hu, Dewen Zeng, Xiaowei Xu, and Yiyu Shi. Semi-supervised contrastive learning for label-efficient medical

image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 481–490. Springer, 2021. 3

[26] Qiangguo Jin, Hui Cui, Changming Sun, Jiangbin Zheng, Leyi Wei, Zhenyu Fang, Zhaopeng Meng, and Ran Su. Semi-supervised histological image segmentation via hierarchical consistency enforcement. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 3–13. Springer, 2022. 2, 6, 7

[27] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020. 2

[28] Jun Li, Yushan Zheng, Kun Wu, Jun Shi, Fengying Xie, and Zhiguo Jiang. Lesion-aware contrastive representation learning for histopathology whole slide images analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 273–282. Springer, 2022. 3

[29] Shuailin Li, Chuyu Zhang, and Xuming He. Shape-aware semi-supervised 3d semantic segmentation for medical images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 552–561. Springer, 2020. 7

[30] Xiaomeng Li, Lequan Yu, Hao Chen, Chi-Wing Fu, Lei Xing, and Pheng-Ann Heng. Transformation-consistent self-ensembling model for semisupervised medical image segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 32(2):523–534, 2020. 7

[31] Xiangde Luo. SSL4MIS. https://github.com/HiLab-git/SSL4MIS, 2020. 6

[32] Xiangde Luo, Jieneng Chen, Tao Song, and Guotai Wang. Semi-supervised medical image segmentation through dual-task consistency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8801–8809, 2021. 1, 7

[33] Xiangde Luo, Minhao Hu, Tao Song, Guotai Wang, and Shaoting Zhang. Semi-supervised medical image segmentation via cross teaching between cnn and transformer. In *Medical Imaging with Deep Learning*, 2021. 1

[34] Xiangde Luo, Wenjun Liao, Jieneng Chen, Tao Song, Yinan Chen, Shichuan Zhang, Nianyong Chen, Guotai Wang, and Shaoting Zhang. Efficient semi-supervised gross target volume of nasopharyngeal carcinoma segmentation via uncertainty rectified pyramid consistency. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 318–329. Springer, 2021. 7

[35] Fei Lyu, Mang Ye, Jonathan Frederik Carlsen, Kenny Erleben, Sune Darkner, and Pong C Yuen. Pseudo-label guided image synthesis for semi-supervised covid-19 pneumonia infection segmentation. *IEEE Transactions on Medical Imaging*, 2022. 2

[36] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. IEEE, 2016. 1

[37] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6707–6717, 2020. 2

[38] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 4

[39] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12674–12684, 2020. 2

[40] Jizong Peng, Ping Wang, Christian Desrosiers, and Marco Pedersoli. Self-paced contrastive learning for semi-supervised medical image segmentation with meta-labels. *Advances in Neural Information Processing Systems*, 34:16686–16699, 2021. 2

[41] 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. 1, 6

[42] Constantin Marc Seibold, Simon Reiß, Jens Kleesiek, and Rainer Stiefelhagen. Reference-guided pseudo-label generation for medical semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2171–2179, 2022. 2

[43] Jiangbo Shi, Tieliang Gong, Chunbao Wang, and Chen Li. Semi-supervised pixel contrastive learning framework for tissue segmentation in histopathological image. *IEEE Journal of Biomedical and Health Informatics*, 2022. 1, 6

[44] Xiaoshuang Shi, Hai Su, Fuyong Xing, Yun Liang, Gang Qu, and Lin Yang. Graph temporal ensembling based semi-supervised convolutional neural network with noisy labels for histopathology image analysis. *Medical image analysis*, 60:101624, 2020. 1

[45] Yinghuan Shi, Jian Zhang, Tong Ling, Jiwen Lu, Yefeng Zheng, Qian Yu, Lei Qi, and Yang Gao. Inconsistency-aware uncertainty estimation for semi-supervised medical image segmentation. *IEEE Transactions on Medical Imaging*, 41(3):608–620, 2021. 7

[46] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020. 2

[47] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017. 2, 5

[48] Vikas Verma, Alex Lamb, Juho Kannala, Yoshua Bengio, and David Lopez-Paz. Interpolation consistency training for semi-supervised learning. In *International Joint Conference on Artificial Intelligence*, pages 3635–3641, 2019. 7

[49] Sulaiman Vesal, Mingxuan Gu, Ronak Kosti, Andreas Maier, and Nishant Ravikumar. Adapt everywhere: unsupervised

adaptation of point-clouds and entropy minimization for multi-modal cardiac image segmentation. *IEEE Transactions on Medical Imaging*, 40(7):1838–1851, 2021. 2

[50] Guotai Wang, Shuwei Zhai, Giovanni Lasio, Baoshe Zhang, Byong Yi, Shifeng Chen, Thomas J Macvittie, Dimitris Metaxas, Jinghao Zhou, and Shaoting Zhang. Semi-supervised segmentation of radiation-induced pulmonary fibrosis from lung ct scans with multi-scale guided dense attention. *IEEE transactions on medical imaging*, 41(3):531–542, 2021. 2

[51] Jianfeng Wang and Thomas Lukasiewicz. Rethinking bayesian deep learning methods for semi-supervised volumetric medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 182–190, 2022. 6, 7

[52] Kaiping Wang, Bo Zhan, Chen Zu, Xi Wu, Jiliu Zhou, Luping Zhou, and Yan Wang. Tripled-uncertainty guided mean teacher model for semi-supervised medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 450–460. Springer, 2021. 2, 7

[53] Kaiping Wang, Bo Zhan, Chen Zu, Xi Wu, Jiliu Zhou, Luping Zhou, and Yan Wang. Semi-supervised medical image segmentation via a tripled-uncertainty guided mean teacher model with contrastive learning. *Medical Image Analysis*, 79:102447, 2022. 2

[54] Tao Wang, Jianglin Lu, Zhihui Lai, Jiajun Wen, and Heng Kong. Uncertainty-guided pixel contrastive learning for semi-supervised medical image segmentation. In *IJCAI*, pages 1444–1450, 2022. 2

[55] Wenguan Wang, Tianfei Zhou, Fisher Yu, Jifeng Dai, Ender Konukoglu, and Luc Van Gool. Exploring cross-image pixel contrast for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7303–7313, 2021. 3

[56] Yixin Wang, Yao Zhang, Jiang Tian, Cheng Zhong, Zhongchao Shi, Yang Zhang, and Zhiqiang He. Double-uncertainty weighted method for semi-supervised learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 542–551. Springer, 2020. 7

[57] Huisi Wu, Zhaoze Wang, Youyi Song, Lin Yang, and Jing Qin. Cross-patch dense contrastive learning for semi-supervised segmentation of cellular nuclei in histopathologic images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11666–11675, 2022. 1, 6

[58] Yicheng Wu, Zongyuan Ge, Donghao Zhang, Minfeng Xu, Lei Zhang, Yong Xia, and Jianfei Cai. Mutual consistency learning for semi-supervised medical image segmentation. *Medical Image Analysis*, 81:102530, 2022. 7

[59] Yicheng Wu, Minfeng Xu, Zongyuan Ge, Jianfei Cai, and Lei Zhang. Semi-supervised left atrium segmentation with mutual consistency training. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 297–306. Springer, 2021. 7

[60] Qingsong Xie, Yuexiang Li, Nanjun He, Munan Ning, Kai Ma, Guoxing Wang, Yong Lian, and Yefeng Zheng. Unsupervised domain adaptation for medical image segmentation by disentanglement learning and self-training. *IEEE Transactions on Medical Imaging*, 2022. 1, 2

[61] Junlin Yang, Nicha C Dvornek, Fan Zhang, Julius Chapiro, MingDe Lin, and James S Duncan. Unsupervised domain adaptation via disentangled representations: Application to cross-modality liver segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 255–263. Springer, 2019. 1

[62] Zhixiong Yang, Junwen Pan, Yanzhan Yang, Xiaozhou Shi, Hong-Yu Zhou, Zhicheng Zhang, and Cheng Bian. Proco: Prototype-aware contrastive learning for long-tailed medical image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 173–182. Springer, 2022. 2

[63] Chenyu You, Weicheng Dai, Fenglin Liu, Haoran Su, Xiaoran Zhang, Lawrence Staib, and James S Duncan. Mine your own anatomy: Revisiting medical image segmentation with extremely limited labels. *arXiv preprint arXiv:2209.13476*, 2022. 6

[64] Chenyu You, Yuan Zhou, Ruihan Zhao, Lawrence Staib, and James S Duncan. Simcvd: Simple contrastive voxel-wise representation distillation for semi-supervised medical image segmentation. *IEEE Transactions on Medical Imaging*, 2022. 1

[65] 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. 2, 7

[66] Shuo Zhang, Jiaojiao Zhang, Biao Tian, Thomas Lukasiewicz, and Zhenghua Xu. Multi-modal contrastive mutual learning and pseudo-label re-learning for semi-supervised medical image segmentation. *Medical Image Analysis*, page 102656, 2022. 2

[67] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. *arXiv preprint arXiv:2010.00747*, 2020. 3

[68] Yuhang Zhang, Xiaopeng Zhang, Jie Li, Robert Qiu, Haohang Xu, and Qi Tian. Semi-supervised contrastive learning with similarity co-calibration. *IEEE Transactions on Multimedia*, 2022. 3

[69] Xinkai Zhao, Chaowei Fang, De-Jun Fan, Xutao Lin, Feng Gao, and Guanbin Li. Cross-level contrastive learning and consistency constraint for semi-supervised medical image segmentation. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2022. 3

[70] Xiangyun Zhao, Raviteja Vemulapalli, Philip Andrew Mansfield, Boqing Gong, Bradley Green, Lior Shapira, and Ying Wu. Contrastive learning for label efficient semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10623–10633, 2021. 8

[71] Ziyuan Zhao, Kaixin Xu, Shumeng Li, Zeng Zeng, and Cuntai Guan. Mt-uda: Towards unsupervised cross-modality

medical image segmentation with limited source labels. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 293–303. Springer, 2021. 1

[72] Mingkai Zheng, Fei Wang, Shan You, Chen Qian, Changshui Zhang, Xiaogang Wang, and Chang Xu. Weakly supervised contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10042–10051, 2021. 3

[73] Zongwei Zhou, Vatsal Sodha, Md Mahfuzur Rahman Siddiquee, Ruibin Feng, Nima Tajbakhsh, Michael B Gotway, and Jianming Liang. Models genesis: Generic autodidactic models for 3d medical image analysis. In *International conference on medical image computing and computer-assisted intervention*, pages 384–393. Springer, 2019. 2