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 to 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 em-
bedding 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 [50, 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 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 \( \mathbb{L}_L \) with its corresponding label set \( \mathbb{Y}_L \) and an unlabeled image set \( \mathbb{L}_U \) 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 \( \mathbb{L}_L \) and \( \mathbb{L}_U \). 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 \( I_i \), 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 \( k = \{1, 2, \ldots, K\} \) 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 positives, and all the patches from other \((K-1)\) classes are negatives. 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:

\[
\text{Avg}_{i,j}^k = \frac{\sum_{m \in P_{i,j}^k} C_i^k(m)}{|P_{i,j}^k|}.
\]  

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, \( \text{Avg}_{i,j}^k \) 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 \( I_i \in \mathbb{L}_L \), the available ground truth \( (Y_i \in \mathbb{Y}_L) \) is used instead of generating its pseudo-label \( Y_i' \).
The embedding of an anchor patch of class \( k \), all the patches with \( n \)-nearest \( \text{Ent}^{k}_{i,j} \) 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 \( \mathbb{U}_L \) (the pseudo-label generation will be explained in subsection 3.4) along with the labeled samples \( (\mathbb{L}_L, \mathbb{Y}_L) \). 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 \( P_v^{\pm} \) if \( v \) is positive (i.e., it has the same pseudo/true class as patch \( P_u \)); otherwise, it is denoted as \( P_v^{\pm} \) (negative key patch of a class different from \( k \)). We denote the embeddings of \( P_u, P_v^{\pm} \) as \( f_u, f_v^{\pm} \) respectively, such that \( \{f_u, f_v^{\pm}, f_v^{-} \} \) \( \rightarrow \mathcal{H}_S(\mathcal{E}_S(\{P_u, P_v^{\pm}, P_v^{-}\})) \). Let \( f_v^{\pm} \sim p(\cdot|k_{\pm}) \) and \( f_v^{-} \sim p(\cdot|k_{-}) \), the expectation of the InfoNCE loss with respect to the joint distribution \( J \), over all the class conditional densities \( p(\cdot|k_{+}) \) and \( p(\cdot|k_{-}) \), is expressed as:

\[
L = -\mathbb{E}_J \log \frac{\exp(f_u^T \cdot f_v^{k_{+}} / \tau)}{\exp(f_u^T \cdot f_v^{k_{+}} / \tau) + \sum_{k_{-}} \sum_{v} \exp(f_u^T \cdot f_v^{k_{-}} / \tau)}
\]

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

\[
L = \mathbb{E}_J \left[ \log \left( \frac{\exp(f_u^T \cdot f_v^{k_{+}} / \tau)}{\exp(f_u^T \cdot f_v^{k_{+}} / \tau) + \sum_{k_{-}} \sum_{v} \exp(f_u^T \cdot f_v^{k_{-}} / \tau)} \right) - \mathbb{E}_p(\cdot|k_{+}) \left( f_u^T \cdot f_v^{k_{+}} / \tau \right) \right]
\]

\[
\leq \log \left[ \mathbb{E}_J \left( \exp(f_u^T \cdot f_v^{k_{+}} / \tau) + \sum_{k_{-}} \sum_{v} \exp(f_u^T \cdot f_v^{k_{-}} / \tau) \right) \right] - f_u^T \mathbb{E}_p(\cdot|k_{+}) (f_v^{k_{+}} / \tau)
\]

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

$$L_u^{PLGCL} = \log \left[ \exp \left( \frac{f^T u \mu_{k+}}{\tau} + \frac{\lambda}{2\tau^2} f^T u \sigma_{k+} f_u \right) \right] + \sum_k \left( \exp \left( \frac{f^T u \mu_{k-}}{\tau} + \frac{\lambda}{2\tau^2} f^T u \sigma_{k-} f_u \right) \right) - f^T u \mu_{k+} / \tau$$

where $\lambda$ is a scaling factor that originates from the term $\sum_{i}$, 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 $\lambda$ is used to scale the effect of $\sigma_{k+}$ that stabilizes the training. The proposed loss $L_u^{PLGCL}$ relies upon reasonable estimation of $\mu_{k+}, \sigma_{k+}, \mu_{k-}, \sigma_{k-}$ from $f^k_{o+}, f^k_{o-}$. 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 $E_S, D_S$, parameterized by $\theta_{E,S}, \theta_{D,S}$, respectively, and the teacher encoder-decoder model $E_T, D_T$, parameterized by $\theta_{E,T}, \theta_{D,T}$. Let the student projection head be denoted as $H_S$, parameterized by $\theta_{H,S}$. 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:

$$L_i^{Reg} = CE \left[ D_S \left( E_S(I_i^w) \right), D_T \left( E_T(I_i^w) \right) \right]$$

where $I_i^w$ and $I_i^s$ 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:

$$L_i^{Sup} = CE \left[ D_S \left( E_S(I_i) \right), Y_i \right]$$

The final objective function is boiled down to:

$$L_i^{total} = \frac{1}{|B_L|} \sum_{I_i \in B_L} L_i^{Sup} + \beta \frac{1}{|B_U|} \sum_{I_i \in B_U} L_i^{Reg} + \gamma \frac{1}{|B|} \sum_{I_i \in B} L_i^{PLGCL}$$

where $B$ is the sampled mini-batch; $B_L, B_U$ 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:

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

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

where $t$ tracks the step number, and $\alpha$ 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 $L^{Reg}$ and $L^{Sup}$ 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 $L^{Reg}$ (refer Equation 6). Additionally, we also compute the supervised CE loss $L^{Sup}$ between the segmentation output of the student model $Y_i^s$ for image $I_i \in \mathbb{I}_L$ and the available ground truth $Y_i \in \mathbb{Y}_L$.

The warm-up training generates initial pseudo-labels, and then the contrastive loss $L^{PLGCL}$ 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.
We experiment with different percentages of labeled data and compare the performance with its counterpart trained 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

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].

4.1. 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].

Figure 2. Visual comparison of segmentation results with different percentages of labeled data for training. 100% indicates fully-supervised setting.
we compare the performance of our work with the existing methods on the MRI dataset. As discussed in section 2, LCLPL [10] performs all the SoTA SemiSL methods like UA-MT [65], URPC [34], DTC [32], MC-Net [59], SASSNet [29] on the 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.

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 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 KitTS19 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-connection. 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.

### 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 $10^7$, $10^8$, $10^9$ pixels in their testing sets, respectively.

<table>
<thead>
<tr>
<th>Method</th>
<th>labeled data (%)</th>
<th>Evaluation Metrics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>DSC</td>
</tr>
<tr>
<td>UA-MT [65]</td>
<td>100%</td>
<td>0.816</td>
</tr>
<tr>
<td>Double-UA [56]</td>
<td>100%</td>
<td>0.883</td>
</tr>
<tr>
<td>MC-Net [59]</td>
<td>100%</td>
<td>0.865</td>
</tr>
<tr>
<td>MC-Net [58]</td>
<td>10%</td>
<td>0.871</td>
</tr>
<tr>
<td>SASSNet [29]</td>
<td>100%</td>
<td>0.835</td>
</tr>
<tr>
<td>DTC [32]</td>
<td></td>
<td>0.827</td>
</tr>
<tr>
<td>LCLPL [10]</td>
<td></td>
<td>0.881</td>
</tr>
<tr>
<td>Ours</td>
<td></td>
<td>0.905</td>
</tr>
<tr>
<td>Supervised</td>
<td>100%</td>
<td>0.923</td>
</tr>
</tbody>
</table>

(a) ACDC

<table>
<thead>
<tr>
<th>Method</th>
<th>labeled data (%)</th>
<th>Evaluation Metrics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>DSC</td>
</tr>
<tr>
<td>UA-MT [65]</td>
<td>20%</td>
<td>0.857</td>
</tr>
<tr>
<td>URPC [34]</td>
<td></td>
<td>0.851</td>
</tr>
<tr>
<td>MC-Net [59]</td>
<td>100%</td>
<td>0.878</td>
</tr>
<tr>
<td>MC-Net [58]</td>
<td>20%</td>
<td>0.885</td>
</tr>
<tr>
<td>SASSNet [29]</td>
<td></td>
<td>0.871</td>
</tr>
<tr>
<td>DTC [32]</td>
<td></td>
<td>0.863</td>
</tr>
<tr>
<td>LCLPL [10]</td>
<td></td>
<td>0.907</td>
</tr>
<tr>
<td>Ours</td>
<td></td>
<td>0.912</td>
</tr>
<tr>
<td>Supervised</td>
<td>100%</td>
<td>0.923</td>
</tr>
</tbody>
</table>

(b) KitTS19

<table>
<thead>
<tr>
<th>Method</th>
<th>labeled data (%)</th>
<th>Evaluation Metrics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>DSC</td>
</tr>
<tr>
<td>UA-MT [65]</td>
<td>100%</td>
<td>0.883</td>
</tr>
<tr>
<td>SASSNet [29]</td>
<td></td>
<td>0.891</td>
</tr>
<tr>
<td>CoraNet [45]</td>
<td></td>
<td>0.896</td>
</tr>
<tr>
<td>DTC [32]</td>
<td>2.5%</td>
<td>0.894</td>
</tr>
<tr>
<td>GBDL [51]</td>
<td>10%</td>
<td>0.911</td>
</tr>
<tr>
<td>Triple-UA [52]</td>
<td>10%</td>
<td>0.887</td>
</tr>
<tr>
<td>Double-UA [56]</td>
<td>10%</td>
<td>0.895</td>
</tr>
<tr>
<td>Ours</td>
<td></td>
<td>0.919</td>
</tr>
<tr>
<td>Supervised</td>
<td>100%</td>
<td>0.934</td>
</tr>
</tbody>
</table>

(c) CRAG

Figure 3. Ablation experiment: t-SNE decomposition of representation space produced by encoders $E_S$ and projection head $H_S$ at different training stages on ACDC dataset (20% labeled) with and without the proposed PLGCL.
decomposition of representation space with and without a small warm-up phase using only PLGCL.

In the pseudo-label generation and refinement, we use a scheme to address the critical issue of class collision in CL. This sufficiently demonstrates the effectiveness of our proposed classes entangle with each other in the feature space. On the other hand, without PLGCL, the embeddings from various other datasets entangle with each other in the feature space, 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 $L^{Sup}$ and $L^{Reg}$ 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 (∼ 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: $\text{Sim}(a, b) = a \cdot b / ||a|| ||b||$. (B) Class Confidence: For a patch $P_{i,j}$, we calculate the average patch confidence $Avg_{i,j}$ (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 $I_{k}^{j}$ 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


[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


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