MINE YOUR OWN ANATOMY: REVISITING MEDICAL IMAGE SEGMENTATION WITH EXTREMELY LIMITED LABELS

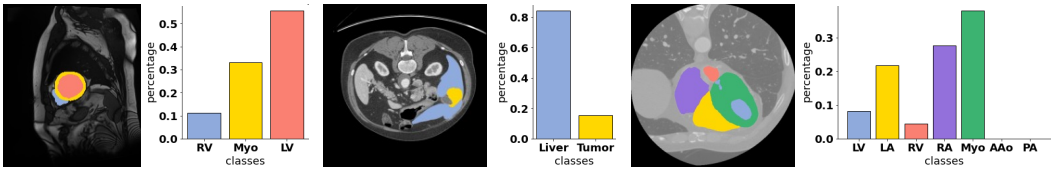
Anonymous authors
Paper under double-blind review

ABSTRACT

Recent studies on contrastive learning have achieved remarkable performance solely by leveraging few labels in the context of medical image segmentation. Existing methods mainly focus on instance discrimination and invariant mapping (*i.e.*, pulling positive samples closer and negative samples apart in the feature space). However, they face three common pitfalls: (1) *tailness*: medical image data usually follows an implicit long-tail class distribution. Blindly leveraging all pixels in training hence can lead to the data imbalance issues, and cause deteriorated performance; (2) *consistency*: it remains unclear whether a segmentation model has learned meaningful and yet consistent anatomical features due to the intra-class variations between different anatomical features; and (3) *diversity*: the intra-slice correlations within the entire dataset have received significantly less attention. This motivates us to seek a principled approach for strategically making use of the dataset itself to discover similar yet distinct samples from *different anatomical views*. In this paper, we introduce a novel semi-supervised medical image segmentation framework termed **M**ine y**O**ur ow**N** **A**natomy (MONA), and make three contributions. First, prior work argues that every pixel equally matters to the model training; we observe empirically that this alone is unlikely to define meaningful anatomical features, mainly due to lacking the supervision signal. We show two simple solutions towards learning invariances – through the use of stronger data augmentations and nearest neighbors. Second, we construct a set of objectives that encourage the model to be capable of decomposing medical images into a collection of anatomical features in an unsupervised manner. Lastly, our extensive results on three benchmark datasets with different labeled settings validate the effectiveness of our proposed MONA which achieves new state-of-the-art under different labeled settings. Perhaps most impressively, MONA trained with 10% labeled – for the first time – outperforms the supervised counterpart on all three datasets. MONA makes minimal assumptions on domain expertise, and hence constitutes a practical and versatile solution in medical image analysis. Codes will be available to public.

1 INTRODUCTION

With the advent of deep learning, medical image segmentation has drawn great attention and substantial research efforts in recent years. Traditional supervised training schemes coupled with large-scale annotated data can engender remarkable performance. However, training with massive high-quality annotated data is infeasible in clinical practice since a large amount of expert-annotated medical data often incurs considerable clinical expertise and time. Under such a setting, this poses the question of how models benefit from a large amount of unlabelled data during training. Recently emerged methods based on contrastive learning (CL) significantly reduce the training cost by learning strong visual representations in an unsupervised manner Wu et al. (2018b); Oord et al. (2018); Hjelm et al. (2019); Chen et al. (2020a); He et al. (2020); Henaff (2020); Misra & Maaten (2020); Hadsell et al. (2006); Grill et al. (2020); Chen et al. (2020b); Caron et al. (2020). A popular way of formulating this idea is through imposing feature consistency to differently augmented views of the same image - which treats each view as an individual instance.



(1) ACDC (2) LiTS (3) MMWHS

Figure 1: Examples of three benchmarks (*i.e.*, ACDC, LiTS, MMWHS) with large intra-class variances.

Despite great promise, the main technical challenges remain: (1) How far is CL from becoming a principled framework for medical image segmentation? (2) Is there any better way to implicitly learn some intrinsic properties from the original data (*i.e.*, the inter-instance relationships and intra-instance invariance)? (3) What will happen if models can only access a few labels in training?

To address the above challenges, we outline three principles below: (1) *tailness*: existing approaches inevitably suffer from class collapse problems – wherein similar pairs from the same latent class are assumed to have the same representation Arora et al. (2019); Chuang et al. (2020); Li et al. (2021). This assumption, however, rarely holds for real-world clinical data. We observe that the long-tail distribution problem has received increasing attention in the computer vision community Kang et al. (2020); Zhu et al. (2014); Cui et al. (2019b); Yang & Xu (2020); Jiang et al. (2021). In contrast, there have been a few prior long-tail works for medical image segmentation. For example, as illustrated in Figure 1, most medical image images follow a Zipf long-tail distribution where various anatomical features share very different class frequencies, which can result in worse performance; (2) *consistency*: considering the scarcity of medical data in practice, augmentations are a widely adopted pre-text task to learn meaningful representations. Intuitively, the anatomical features should be semantically consistent across different transformations and deformations. Thus, it is important to assess whether the model is robust to diverse views of anatomy; (3) *diversity*: recent work Zheng et al. (2021); Azabou et al. (2021); Van Gansbeke et al. (2021) pointed out that going beyond simple augmentations to create more diverse views can learn more discriminative anatomical features. At the same time, this is particularly challenging to both introduce sufficient diversity and preserve the anatomy of the original data, especially in data-scarce clinical scenarios. To deploy into the wild, we need to quantify and address three research gaps from *different anatomical views*.

In this paper, we present **M**ine y**O**ur ow**N** **A**natomy (MONA), a novel contrastive semi-supervised medical segmentation framework, based on different anatomical views. The workflow of MONA is illustrated in Figure 2. The **key innovation** in MONA is to seek diverse views (*i.e.*, augmented/mined views) of different samples whose anatomical features are *homogeneous* within the *same class type*, while *distinctive* for *different class types*. We make the following contributions. First, we consider the problem of *tailness*. An issue is that label classes within medical images typically exhibit a long-tail distribution. Another one, technically more challenging, is the fact that there is only a few labeled data and large quantities of unlabeled ones during training. Intuitively we would like to sample more pixel-level representations from tail classes. Thus, we go beyond the naïve setting of instance discrimination in CL Chen et al. (2020a); He et al. (2020); Grill et al. (2020) by decomposing images into diverse and yet consistent anatomical features, each belonging to different classes. In particular, we propose to use pseudo labeling and knowledge distillation to learn better pixel-level representations within multiple semantic classes in a training mini-batch. Considering performing pixel-level CL with medical images is impractical for both memory cost and training time, we then adopt active sampling strategies Liu et al. (2021) such as in-batch hard negative pixels, to better discriminate the representations at a larger scale.

We further address the two other challenges: *consistency* and *diversity*. The success of the common CL theme is mainly attributed to invariant mapping Hadsell et al. (2006) and instance discrimination Wu et al. (2018b); Chen et al. (2020a). Starting from these two key aspects, we try to further improve the segmentation quality. More specifically, we suggest that *consistency* to transformation (equivariance) is an effective strategy to establish the invariances (*i.e.*, anatomical features and shape variance) to various image transformations. Furthermore, we investigate two ways to include diversity-promoting views in sample generation. First, we incorporate a memory buffer to alleviate the demand for large batch size, enabling much more efficient training without inhibiting segmentation quality. Second,

we leverage stronger augmentations and nearest neighbors to mine views as positive views for more semantic similar contexts.

Extensive experiments are conducted on a variety of datasets and the latest CL frameworks (*i.e.*, MoCo He et al. (2020), SimCLR Chen et al. (2020a), BYOL Grill et al. (2020), and ISD Tejankar et al. (2021)), which consistently demonstrate the effectiveness of our proposed MONA. For example, our MONA establishes the **new state-of-the-art** performance, compared to both the state-of-the-art semi-supervised and fully-supervised approaches with 10% label ratio. We also present a systematic evaluation for analyzing why our approach performs so well and how different factors contribute to the final performance. We hope our findings will provide useful insights on medical image segmentation to other researchers.

2 RELATED WORK

Medical image segmentation Medical image segmentation aims to assign a class label to each pixel in an image, and plays a major role in real-world applications, such as assisting the radiologists for better disease diagnosis and reduced cost. With sufficient annotated training data, significant progress has been achieved with the introduction of Fully convolutional networks (FCN) Long et al. (2015) and UNet Ronneberger et al. (2015). Follow-up works can be categorized into two main directions. One direction is to improve modern segmentation network design. Many CNN-based Simonyan & Zisserman (2014); He et al. (2016) and Transformer-like Vaswani et al. (2017); Dosovitskiy et al. (2020) model variants Milletari et al. (2016); Chen et al. (2017); Alom et al. (2018); Oktay et al. (2018); Chen et al. (2018); Wu et al. (2018a; 2019); Chen et al. (2021a); Cao et al. (2021); Xie et al. (2021); Hatamizadeh et al. (2021); Valanarasu et al. (2021); Desai et al. (2021); Xu et al. (2021); Xu et al.; Isensee et al. (2021); You et al. (2022a) have been proposed since then. For example, some works Chen et al. (2017; 2018); Dai et al. (2017) proposed to use dilated/atrous/deformable convolutions with larger receptive fields for more dense anatomical features. Other works Chen et al. (2021a); Cao et al. (2021); Xie et al. (2021); Hatamizadeh et al. (2021); Valanarasu et al. (2021); You et al. (2022a) include Transformer blocks to capture more long-range information, achieving the impressive performance. A parallel direction is to select proper optimization strategies, by designing loss functions to learn meaningful representations Lin et al. (2017b); Xue et al. (2019); Shi et al. (2021). However, those methods assume access to a large, labeled dataset. This restrictive assumption makes it challenging to deploy in most real-world clinical practices. In contrast, our MONA is more robust as it leverages only a few labeled data and large quantities of unlabeled one in the learning stage.

Semi-supervised learning (SSL) The goal in robust SSL is to improve the medical segmentation performance by taking advantage of large amounts of unlabelled data during training. It can be roughly categorized into three groups: (1) self-training by generating unreliable pseudo-labels for performance gains, such as pseudo-label estimation Lee et al. (2013); Bai et al. (2017); Fan et al. (2020); Chen et al. (2021b), model uncertainty Yu et al. (2019); Graham et al. (2019); Jungo & Reyes (2019); Mehrtash et al. (2020); Zeng et al. (2019); Nair et al. (2020); Camarasa et al. (2020); Cao et al. (2020), confidence estimation Blundell et al. (2015); Gal & Ghahramani (2016); Kendall & Gal (2017), and noisy student Xie et al. (2020a); (2) consistency regularization Bortsova et al. (2019); Cui et al. (2019a); Zhou et al. (2020); Fotedar et al. (2020); Fang & Li (2020) by integrating consistency corresponding to different transformation, such as pi-model Sajjadi et al. (2016), co-training Qiao et al. (2018); Zhou et al. (2019), and mean-teacher Tarvainen & Valpola (2017); Li et al. (2020b); Reiß et al. (2021); (3) other training strategies such as adversarial training Zhang et al. (2017); Nie et al. (2018); Zhang et al. (2018); Zheng et al. (2019); Li et al. (2020a); Valvano et al. (2021) and entropy minimization Grandvalet & Bengio (2004). In contrast to these works, we do not explore more advanced pseudo-labelling strategy to learn spatially structured representations. In this work, we are the first to explore a novel direction for discovering distinctive and semantically consistent anatomical features without image-level or region-level labels. Further, we expect that our findings can be relevant for other medical image segmentation frameworks.

Contrastive learning CL has recently emerged as a promising paradigm for medical image segmentation via exploiting abundant unlabeled data, leading to state-of-the-art results Chaitanya et al. (2020); Xie et al. (2020b); You et al. (2021); Chaitanya et al. (2021); Hu et al. (2021); Xiang et al. (2021); Zeng et al. (2021); You et al. (2022b). The high-level idea of CL is to pull closer the different

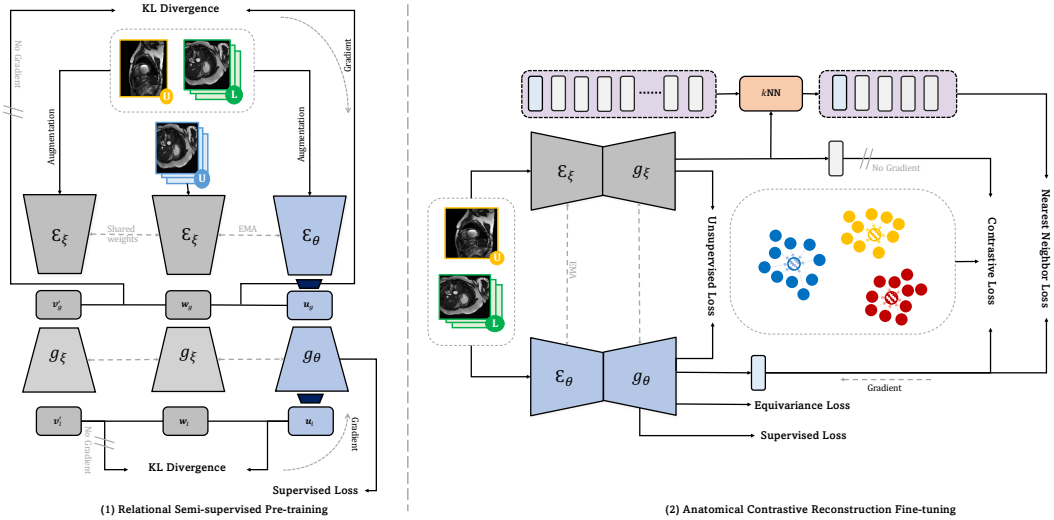


Figure 2: Overview of the MONA framework including two stages: (1) relational semi-supervised pre-training, (2) our proposed anatomical contrastive reconstruction fine-tuning. Note that U and L denote unlabeled and labeled data.

augmented views of the same instance but pushes apart all the other instances away. Intuitively, differently augmented views of the same image are considered *positives*, while all the other images serve as *negatives*. The major difference between different CL-based frameworks lies in the augmentation strategies to obtain *positives* and *negatives*. A few very recent studies Kang et al. (2020); Jiang et al. (2021) confirm the superiority of CL of addressing imbalance issues in image classification. Moreover, existing CL frameworks Chaitanya et al. (2020); You et al. (2021); Hu et al. (2021) mainly focus on the instance level discrimination (*i.e.*, different augmented views of the same instance should have similar anatomical features or clustered around the class weights). However, we argue that not all negative samples equally matter, and the above issues have not been explored from the perspective of medical image segmentation, considering the class distributions in the medical image are perspectives diverse and always exhibit long tails Galdran et al. (2021); Roy et al. (2022). Inspired by the aforementioned, we address these two issues in medical image segmentation - two appealing perspectives that still remain under-explored.

3 MINE YOUR OWN ANATOMY (MONA)

3.1 FRAMEWORK

Overview. We introduce our contrastive learning framework (See Figure 2), which includes (1) relational semi-supervised pre-training, and (2) anatomical contrastive reconstruction fine-tuning. The key idea is to seek diverse yet semantically consistent views whose anatomical features are *homogeneous* within the *same class type*, while *distinctive* for *different class types*. In this paper, our pre-training stage is built upon ISD Tejankar et al. (2021) - a competitive framework for image classification. The *main differences* between ISD and MONA are: MONA is more tailored to medical image segmentation, *i.e.*, considering the dense nature of this problem both in global and local manner, and can generalize well to those long-tail scenarios. Also, our principles are expected to apply to other CL framework ((*i.e.*, MoCo He et al. (2020), SimCLR Chen et al. (2020a), BYOL Grill et al. (2020)). More detailed analysis can be found in the Appendix C.

Pre-training preliminary. Let (X, Y) be our dataset, including training images $\mathbf{x} \in X$ and their corresponding $\mathcal{C}$-class segmentation labels $\mathbf{y} \in Y$, where X is composed of N labeled and M unlabeled slices. Note that, for brevity, $\mathbf{y}$ can be either sampled from Y or pseudo-labels. The *student* and *teacher* networks $\mathcal{F}$, parameterized by weights θ and ξ, each consist of a encoder $\mathcal{E}$ and a decoder $\mathcal{D}$ (*i.e.*, UNet Ronneberger et al. (2015)). Concretely, given a sample $\mathbf{s}$ from our unlabeled dataset, we have two ways to generate views: (1) we formulate *augmented* views (*i.e.*, $\mathbf{x}, \mathbf{x}'$) through two different augmentation chains; and (2) we create d *mined* views (*i.e.*, $\mathbf{x}_{r,i}$) by randomly selecting

from the unlabeled dataset followed by additional augmentation.[1] We then fed the *augmented* views to both $\mathcal{F}_\theta$ and $\mathcal{F}_\xi$, and the *mined* views to $\mathcal{F}_\xi$. Similar to Chaitanya et al. (2020), we adopt the global and local instance discrimination strategies in the latent and output feature spaces.[2] Specifically, the encoders generate global features $\mathbf{z}_g = \mathbb{E}_\theta(\mathbf{x})$, $\mathbf{z}'_g = \mathbb{E}_\xi(\mathbf{x}')$, and $\mathbf{z}_{r,g} = \mathbb{E}_\xi(\mathbf{x}_r)$, which are then fed into the nonlinear projection heads to obtain $\boldsymbol{v}_g = h_\theta(\mathbf{z}_g)$, $\boldsymbol{v}'_g = h_\xi(\mathbf{z}'_g)$, and $\mathbf{w}_g = h_\xi(\mathbf{z}_{r,g})$. The *augmented* embeddings from the *student* network are further projected into secondary space, *i.e.*, $\mathbf{u}_g = h'_\theta(\boldsymbol{v}_g)$. We calculate similarities across *mined* views and *augmented* views from the *student* and *teacher* in both global and local manners. Then a `softmax` function is applied to process the calculated similarities, which models the relationship distributions:

$$\mathbf{s}_\theta = \log \frac{\exp\big(\mathrm{sim}\big(\mathbf{u}, \mathbf{w}\big)/\tau_\theta\big)}{\sum_{j=1}^{k} \exp\big(\mathrm{sim}\big(\mathbf{u}, \mathbf{w}_j\big)/\tau_\theta\big)}, \quad \mathbf{s}_\xi = \log \frac{\exp\big(\mathrm{sim}\big(\boldsymbol{v}', \mathbf{w}\big)/\tau_\xi\big)}{\sum_{j=1}^{k} \exp\big(\mathrm{sim}\big(\boldsymbol{v}', \mathbf{w}_j\big)/\tau_\xi\big)}, \tag{1}$$

where τ_θ and τ_ξ are different temperature parameters, and $\mathrm{sim}(\cdot,\cdot)$ denotes cosine similarity. The unsupervised instance discrimination loss (*i.e.*, Kullback-Leibler divergence $\mathcal{KL}$) can be defined as:

$$\mathcal{L}_{\text{inst}} = \mathcal{KL}(\mathbf{s}_\theta || \mathbf{s}_\xi). \tag{2}$$

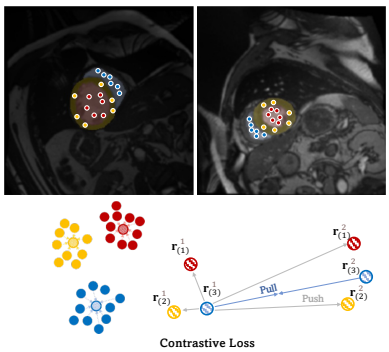
The parameters ξ of $\mathcal{F}_\xi$ is updated as: $\xi = t\xi + (1-t)\theta$ with $t = 0.99$ as a momentum hyperparameter. In our pre-training stage, the total loss is the sum of global and local instance discrimination loss $\mathcal{L}_{\text{inst}}$ (on pseudo-labels), and supervised segmentation loss $\mathcal{L}_{\text{sup}}$ (*i.e.*, equal combination of dice loss and cross-entropy loss on ground-truth labels): $\mathcal{L}_{\text{inst}}^{\text{gloabl}} + \mathcal{L}_{\text{inst}}^{\text{local}} + \mathcal{L}_{\text{sup}}$.

Principles. As shown in Figure 2, the principles behind `MONA` (*i.e.*, the second anatomical contrastive reconstruction stage) aim to ensure tailness, consistency, and diversity. Concretely, tailness is for actively sampling more tail class hard pixels; consistency ensures the feature invariances; and diversity further encourages to discover more anatomical features in different images.

3.2 Anatomical Contrastive Reconstruction

Tailness. Motivated by the observations (Figure 1), our primary cue is that medical images naturally exhibit an imbalanced or long-tailed class distribution, wherein many class labels are associated with only a few pixels. To generalize well on such *imbalanced* setting, we propose to use *anatomical contrastive formulation* (**ACF**).

Here we additionally attach the representation heads to fuse the multi-scale features with the feature pyramid network (`FPN`) Lin et al. (2017a) structure and generate the m-dimensional representations with consecutive convolutional layers. The high-level idea is that the features should be very *similar* among the same class type, while very *dissimilar* across different class types. Particularly for long-tail medical data, a naïve application of this idea would require substantially computational resources proportional to the square of the number of pixels within the dataset, and naturally overemphasize the anatomy-rich head classes and leaves the tail classes under-learned in learning invariances, both of which suffer performance drops.



To this end, we address this issue by actively sampling a set of pixel-level anchor representations $\mathrm{r}_q \in \mathcal{R}_q^c$ (*queries*), pulling them closer to the class-averaged mean of representations $\mathrm{r}_k^{c,+}$ within this class c (*positive keys*), and pushing away from representations $\mathrm{r}_k^- \in \mathcal{R}_k^c$ from other classes (*negative keys*). Formally, the contrastive loss is defined as:

$$\mathcal{L}_{\text{contrast}} = \sum_{c \in \mathcal{C}} \sum_{\mathrm{r}_q \sim \mathcal{R}_q^c} -\log \frac{\exp(\mathrm{r}_q \cdot \mathrm{r}_k^{c,+}/\tau)}{\exp(\mathrm{r}_q \cdot \mathrm{r}_k^{c,+}/\tau) + \sum_{\mathrm{r}_k^- \sim \mathcal{R}_k^c} \exp(\mathrm{r}_q \cdot \mathrm{r}_k^-/\tau)}, \tag{3}$$

[1] Note that the subscript i is omitted for simplicity in following contexts.

[2] Here we omit details of local instance discrimination strategy for simplicity because the global and local instance discrimination experimental setups are similar.

where $\mathcal{C}$ denotes a set of all available classes for each mini-batch, and τ is a temperature hyperparameter. Suppose $\mathcal{A}$ is a collection including all pixel coordinates within $\mathbf{x}$, these representations are:

$$\mathcal{R}_q^c = \bigcup_{[m,n]\in\mathcal{A}} \mathbb{1}(\mathbf{y}_{[m,n]}=c)\,\mathrm{r}_{[m,n]},\ \mathcal{R}_k^c = \bigcup_{[m,n]\in\mathcal{A}} \mathbb{1}(\mathbf{y}_{[m,n]}\neq c)\,\mathrm{r}_{[m,n]},\ \mathrm{r}_k^{c,+} = \frac{1}{|\mathcal{R}_q^c|}\sum_{\mathrm{r}_q\in\mathcal{R}_q^c}\mathrm{r}_q\,. \quad (4)$$

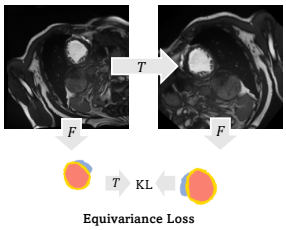
We then note that CL might benefit more, where the instance discrimination task is achieved by incorporating more positive and negative pairs. However, naively unrolling CL to this setting is impractical since it requires extra memory overheads that grow proportionally with the amount of instance discrimination tasks. To this end, we adopt a random set (*i.e.*, the mini-batch) of other images. Intuitively, we would like to maximize the anatomical similarity between all the representations from the query class, and analogously minimize all other class representations. We then create a graph $\mathcal{G}$ to compute the pair-wise class relationship: $\mathcal{G}[p,q] = \left(\mathrm{r}_k^{p,+}\cdot\mathrm{r}_k^{q,+}\right), \forall p,q\in\mathcal{C},$ and $p\neq q$, where $\mathcal{G}\in\mathbb{R}^{|\mathcal{C}|\times|\mathcal{C}|}$. Here finding the accurate decision boundary can be formulated mathematically by normalizing the pair-wise relationships among all negative class representations via the `softmax` operator. To address the challenge in *imbalanced* medical image data, we define the pseudo-label (*i.e.*, easy and hard queries) based on a defined threshold as follows:

$$\mathcal{R}_q^{c,\,\text{easy}} = \bigcup_{\mathrm{r}_q\in\mathcal{R}_q^c} \mathbb{1}(\hat{\mathbf{y}}_q > \delta_\theta)\mathrm{r}_q, \quad \mathcal{R}_q^{c,\,\text{hard}} = \bigcup_{\mathrm{r}_q\in\mathcal{R}_q^c} \mathbb{1}(\hat{\mathbf{y}}_q \leq \delta_\theta)\mathrm{r}_q, \quad (5)$$

where $\hat{\mathbf{y}}_q$ is the c^{th}-class pseudo-label corresponding to r_q, and δ_θ is the user-defined threshold. For further improvement in long-tail scenarios, we construct a class-aware memory bank He et al. (2020) to store a fixed number of negative samples per class c.

Consistency.

The proposed ACF is designed to address *imbalanced* issues, but *anatomical consistency* remains to be weak in the long-tail medical image setting since medical segmentation should be robust to different tissue types which show different anatomical variations. We hence construct a random image transformation $\mathcal{T}$ and define the equivariance loss on both labeled and unlabeled data by measuring the feature consistency distance between each original segmentation map and the segmentation map generated from the transformed image:



$$\mathcal{L}_{\text{eqv}}(\mathbf{x},\mathcal{T}(\mathbf{x})) = \sum_{\mathbf{x}\in X} \mathcal{KL}\left(\mathcal{T}(\mathcal{F}_\theta(\mathbf{x})),\mathcal{F}_\theta(\mathcal{T}(\mathbf{x}))\right) + \mathcal{KL}\left(\mathcal{F}_\theta(\mathcal{T}(\mathbf{x})),\mathcal{T}(\mathcal{F}_\theta(\mathbf{x}))\right)\,. \quad (6)$$

Here we define $\mathcal{T}$ on both the input image $\mathbf{x}$ and $\mathcal{F}_\theta(\mathbf{x})$, via the random transformations (*i.e.*, affine, intensity, and photo-metric augmentations), since the model should learn to be robust and invariant to these transformations.

Diversity. Oversampling too many images from the random set would create extra memory overhead, and more importantly, our finding also uncovers that a large number of random images might not necessarily help impose additional invariances between neighboring samples since redundant images might introduce additional noise during training (see the Appendix D). Therefore, we formulate our insight as an auxiliary loss that regularizes the representations - keeping the anatomical contrastive reconstruction task as the main force. In practice, we first search for K-nearest neighbors from the first-in-first-out (FIFO) memory bank He et al. (2020), and then use the nearest neighbor loss $\mathcal{L}_{\text{nn}}$ based on the Mean Squared Error (MSE), to exploit the inter-instance relationship.

Setup. The total loss $\mathcal{L}_{\text{total}}$ is the sum of contrastive loss $\mathcal{L}_{\text{contrast}}$ (on both ground-truth labels and pseudo-labels), equivariance loss $\mathcal{L}_{\text{eqv}}$ (on both ground-truth labels and pseudo-labels), nearest neighbors loss $\mathcal{L}_{\text{nn}}$ (on both ground-truth labels and pseudo-labels), unsupervised cross-entropy loss $\mathcal{L}_{\text{unsup}}$ (on pseudo-labels) and supervised segmentation loss $\mathcal{L}_{\text{sup}}$ (on ground-truth labels): $\mathcal{L}_{\text{sup}} + \lambda_1\mathcal{L}_{\text{contrast}} + \lambda_2\mathcal{L}_{\text{eqv}} + \lambda_3\mathcal{L}_{\text{unsup}} + \lambda_4\mathcal{L}_{\text{nn}}$. See the Appendix D for an ablation study of hyperparameters.

Table 1: Comparison of segmentation performance (DSC[%]/ASD[mm]) on ACDC and LiTS under three labeled ratio settings (1%, 5%, 10%). The best results are indicated in **bold**.

Method	ACDC						LiTS					
	1% Labeled		5% Labeled		10% Labeled		1% Labeled		5% Labeled		10% Labeled	
	DSC ↑	ASD ↓	DSC ↑	ASD ↓	DSC ↑	ASD ↓	DSC ↑	ASD ↓	DSC ↑	ASD ↓	DSC ↑	ASD ↓
UNet-F Ronneberger et al. (2015)	89.9	0.621	89.9	0.621	89.9	0.621	68.2	16.9	68.2	16.9	68.2	16.9
UNet-L	14.5	19.3	51.7	13.1	74.4	2.20	57.0	34.6	60.4	30.4	61.6	28.3
EM Vu et al. (2019)	21.1	21.4	59.8	5.64	75.7	2.73	56.6	38.4	61.2	33.3	62.9	38.5
CCT Ouali et al. (2020)	30.9	28.2	59.1	10.1	75.9	3.60	52.4	52.3	60.6	48.7	63.8	31.2
DAN Zhang et al. (2017)	34.7	25.7	56.4	15.1	76.5	3.01	57.2	27.1	62.3	25.8	63.2	30.7
URPC Luo et al. (2021)	32.2	26.9	58.9	8.14	73.2	2.68	55.5	34.6	62.4	37.8	63.0	43.1
DCT Qiao et al. (2018)	36.0	24.2	58.5	10.8	78.1	2.64	57.6	38.5	60.8	34.4	61.9	31.7
ICT Verma et al. (2019)	35.8	21.3	59.0	4.59	75.1	0.898	58.3	32.2	60.1	39.1	62.5	32.4
MT Tarvainen & Valpola (2017)	36.8	19.6	58.3	11.2	80.1	2.33	56.7	34.3	61.9	40.0	63.3	26.2
UAMT Yu et al. (2019)	35.2	24.3	61.0	7.03	77.6	3.15	57.8	41.9	61.0	47.0	62.3	26.0
CPS Chen et al. (2021b)	37.1	30.0	61.0	2.92	78.8	3.41	57.7	39.6	62.1	36.0	64.0	23.6
GCL Chaitanya et al. (2020)	59.7	14.3	70.6	2.24	87.0	**0.751**	59.3	29.5	63.3	20.1	65.0	37.2
SCS Hu et al. (2021)	59.4	12.7	73.6	5.37	84.2	2.01	57.8	39.6	61.5	28.8	64.6	33.9
PLC Chaitanya et al. (2021)	58.8	15.1	70.6	2.67	87.3	1.34	56.6	41.6	62.7	26.1	68.2	**16.9**
• MONA (ours)	**82.6**	**2.03**	**88.8**	**0.62**	**90.7**	0.864	**64.1**	**20.9**	**67.3**	**16.4**	**69.3**	18.0

4 EXPERIMENTS

In this section, we evaluate our proposed MONA on three popular medical image segmentation datasets under varying labeled ratio settings: the ACDC dataset Bernard et al. (2018), the LiTS dataset Bilic et al. (2019), and the MMWHS dataset Zhuang & Shen (2016) (See Appendix B). Moreover, to further validate our approach's unsupervised imbalance handling ability, we consider a more realistic and more challenging scenario, wherein the models would only have access to the extremely limited labeled data (*i.e.*, 1% labeled ratio) and large quantities of unlabeled one in training. For all experiments, we follow the same training and testing protocol. See the Appendix A for more implementation details used in the experiments.

4.1 MAIN RESULTS

We show the effectiveness of our method under three different label ratios (*i.e.*, 1%, 5%, 10%). We also compare MONA with various state-of-the-art SSL and fully-supervised methods on three datasets: ACDC Bernard et al. (2018), LiTS Bilic et al. (2019), MMWHS Zhuang & Shen (2016). We choose 2D UNet Ronneberger et al. (2015) as backbone, and compare against SSL methods including UNet trained with full/limited supervisions (UNet-F/UNet-L), EM Vu et al. (2019), CCT Ouali et al. (2020), DAN Zhang et al. (2017), URPC Luo et al. (2021), DCT Qiao et al. (2018), ICT Verma et al. (2019), MT Tarvainen & Valpola (2017), UAMT Yu et al. (2019), CPS Chen et al. (2021b), SCS Hu et al. (2021), GCL Chaitanya et al. (2020), and PLC Chaitanya et al. (2021). We report quantitative comparisons on ACDC and LiTS in Table 1, and average all our results over three independent runs. (More results on MMWHS in the Appendix B.)

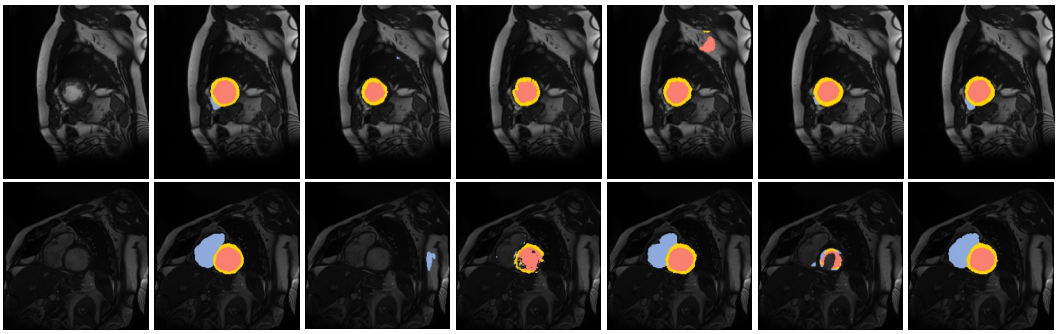


Figure 3: Visualization of segmentation results on ACDC with 5% label ratio. As is shown, MONA consistently yields more accurate predictions and better boundary adherence compared to all other SSL methods. Different anatomical classes are shown in different colors (RV: ; Myo: ; LV:).

ACDC. We benchmark performances on ACDC with respect to different amounts of labeled ratios (*i.e.*, 1%, 5%, 10%). The following observations can be drawn: *First*, our proposed MONA significantly

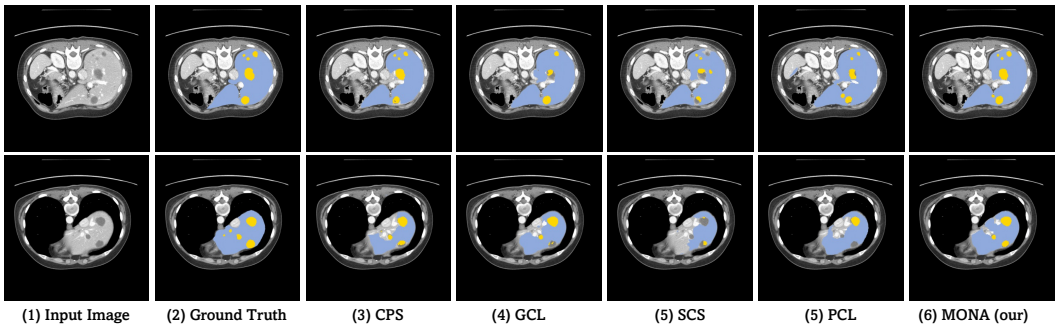


Figure 4: Visualization of segmentation results on LiTS with 5% labeled ratio. As is shown, MONA consistently produces sharp and accurate object boundaries compared to all other SSL methods. Different anatomical classes are shown in different colors (Liver: ; Tumor:).

outperforms all other SSL methods under three different label ratios. Especially, with only extremely limited labeled data available (*e.g.*, 1%), our method obtains massive gains of 82.6% and 2.03 in Dice and ASD (*i.e.*, dramatically improving the performance from 59.4% to 82.6%). *Second*, our method achieves consistently improved performance, and performs better or on par with the fully-supervised approach under all three different label ratios. In particular, MONA with limited labeled training data available (e.g., 10%) – **for the first time** – surpasses the fully supervised counterparts. For example, the best Dice score on ACDC rises from 89.9% to 90.7%. *Third*, as shown in Figure 3, we can see the clear advantage of MONA, where the anatomical boundaries of different tissues are clearly more pronounced such as RV and Myo regions. As seen, our method is capable of producing consistently sharp and accurate object boundaries across various challenge scenarios.

LiTS. We then evaluate MONA on LiTS, using 1%, 5%, 10% labeled ratios. The results are summarized in Table 1 and Figure 4. The conclusions we can draw are highly consistent with the above ACDC case: *First*, at the different label ratios (*i.e.*, 1%, 5%, 10%), MONA consistently outperforms all the other SSL methods, which again demonstrates the effectiveness of learning representations for the inter-class correlations and intra-class invariances under imbalanced class-distribution scenarios. In particular, our MONA, trained on a 1% labeled ratio (*i.e.*, extremely limited labels), dramatically improves the previous best averaged Dice score from 59.3% to 64.1% by a large margin, and even performs on par with previous SSL methods using 10% labeled ratio. *Second*, the most impressive results come from MONA at 10% label ratio. To the best of our knowledge, this is the first time in the literature that SSL schemes trained at 10% label ratio outperform the fully-supervised model by 1.1% improvements in Dice (*i.e.*, from 68.2% to 69.3%). *Third*, as shown in Figure 4, we observe that MONA is able to produce more accurate results compared to the previous best schemes.

Overall, we conclude that MONA provides robust performance on all the medical datasets we evaluated, exceeding that of the fully-supervised baseline, and outperforming all other SSL methods.

4.2 Ablation Study

In this subsection, we conduct comprehensive analyses to understand the inner workings of MONA on ACDC under 5% labeled ratio. Note that for reproducibility, we report the average performance of three independent runs with different random seeds. More results and details about our case study are referred to the Appendix C and D.

Effects of Different Components. Our key observation is that it is crucial to build meaningful anatomical representations for the inter-class correlations and intra-class invariances under imbalanced class-distribution scenarios can further improve performance. Upon our choice of architecture, we first consider a naïve baseline (ISD). To validate this, we experiment with the key components in MONA on ACDC, including: (1) tailness, (2) consistency, and (3) diversity. The results are in Table 2. As is shown, each key component makes a clear difference and leveraging all of them contributes to the remarkable performance improvements. This suggests the importance of learning meaningful representations for the inter-class correlations and intra-class invariances within the entire dataset. The intuitions behind each concept are as follows: (1) **Only *tailness:*** many anatomy-rich head classes would be sampled; (2) **Only *consistency:*** it would lead to object collapsing due to the different anatomical variations; (3) **Only *diversity:*** oversampling too many negative samples often comes at

the cost of performance degradation. By combining *tailness*, *consistency*, and *diversity*, our method confers a significant advantage at representation learning in imbalanced feature similarity, semantic consistency and anatomical diversity, which further highlights the superiority of our proposed MONA.

Table 2: Ablation on model component: (1) tailness; (2) consistency; (3) diversity, compared to the Vanilla and our proposed MONA.

Method	Metrics	
	Dice[%] ↑	ASD[mm] ↓
Vanilla	67.4	6.53
w/ tailness	87.1	1.02
w/ consistency	74.1	11.8
w/ diversity	74.3	10.9
w/ tailness + consistency	88.1	0.864
w/ consistency + diversity	80.2	6.11
w/ tailness + diversity	88.0	1.13
• MONA (ours)	**88.8**	**0.62**

Table 3: Ablation on augmentation strategies for MONA on the ACDC and LiTS dataset under 5% labeled ratio.

Dataset	Student Aug.	Teacher Aug.	Metrics Dice[%] ↑	Metrics ASD[mm] ↓
ACDC	Weak	Weak	86.0	1.02
	Weak	Strong	88.8	0.62
	Strong	Weak	86.4	2.83
	Strong	Strong	88.8	2.07
LiTS	Weak	Weak	62.3	26.5
	Weak	Strong	67.3	16.4
	Strong	Weak	64.3	34.7
	Strong	Strong	66.5	21.1

Effects of Different Augmentations. In addition to further improving the quality and stability in anatomical representation learning, we claim that MONA also gains robustness using augmentation strategies. For augmentation strategies, previous works Tejankar et al. (2021); Zheng et al. (2021); Sohn et al. (2020) show that composing the weak augmentation strategy for the "pivot-to-target" model (*i.e.*, trained with limited labeled data and a large number of unlabeled data) is helpful for anatomical representation learning since the standard contrastive strategy is too aggressive, intuitively leading to a "hard" task (*i.e.*, introducing too many disturbances and yielding model collapses). Here we examine whether and how applying different data augmentations helps MONA. In this work, we implement the weak augmentation to the student's input as random rotation, random cropping, horizontal flipping, and strong augmentation to the teacher's input as random rotation, random cropping, horizontal flipping, random contrast, CutMix French et al. (2020), brightness changes Perez et al. (2018), morphological changes (diffeomorphic deformations). We summarize the results in Table 3, and list the following observations: (1) **weak augmentations benefits more**: composing the weak augmentation for the student model and strong augmentation for the teacher model significantly boosts the performance across two benchmark datasets. (2) **same augmentation pairs do not make more gains**: interestingly, applying same type of augmentation pairs does not lead to the best performance compared to different types of augmentation pairs. We postulate that composing different augmentations can be considered as a harder albeit more useful strategy for anatomical representation learning, making feature more generalizable.

5 Conclusion and Discussion of Broader Impact

In this paper, we have presented MONA, a semi-supervised contrastive learning method for medical image segmentation. We start from the observations that medical image data always exhibit a long-tail class distribution, and the same anatomical objects (*i.e.*, liver regions for two people) are more similar to each other than different objects (*e.g.*liver and tumor regions). We further expand upon this idea by introducing anatomical contrastive formulation, as well as equivariance and invariances constraints. An extensive empirical study shows that we can formulate a generic set of perspectives that allows us to learn meaningful representations across different anatomical features, which can dramatically improve the segmentation quality and alleviate the training memory bottleneck. Extensive experiments on three datasets demonstrate the state-of-the-art performance of our proposed framework in the long-tailed medical data regimes with extremely limited labels. We believe our results contribute to a better understanding of medical image segmentation and point to new avenues for long-tailed medical image data in realistic clinical applications.

Broader Impact. This research aims to study and improve contrastive learning methods for learning useful representations with only extremely limited labels in the long-tail medical data regimes. Our findings show that our proposed framework can potentially benefit the effectiveness of anatomical representation learning and reduce computational costs, leading to realistic deployments in a large variety of real-world clinical applications. Besides, we should address the challenges of fairness or privacy in medical imaging domain as our future research direction.

REFERENCES

Md Zahangir Alom, Mahmudul Hasan, Chris Yakopcic, Tarek M Taha, and Vijayan K Asari. Recurrent residual convolutional neural network based on U-Net (R2U-Net) for medical image segmentation. *arXiv preprint arXiv:1802.06955*, 2018.

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.

Mehdi Azabou, Mohammad Gheshlaghi Azar, Ran Liu, Chi-Heng Lin, Erik C Johnson, Kiran Bhaskaran-Nair, Max Dabagia, Bernardo Avila-Pires, Lindsey Kitchell, Keith B Hengen, et al. Mine your own view: Self-supervised learning through across-sample prediction. *arXiv preprint arXiv:2102.10106*, 2021.

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 (MICCAI)*, 2017.

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

Patrick Bilic, Patrick Ferdinand Christ, Eugene Vorontsov, Grzegorz Chlebus, Hao Chen, Qi Dou, Chi-Wing Fu, Xiao Han, Pheng-Ann Heng, Jürgen Hesser, et al. The liver tumor segmentation benchmark (lits). *arXiv preprint arXiv:1901.04056*, 2019.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International Conference on Machine Learning (ICML)*, 2015.

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

Robin Camarasa, Daniel Bos, Jeroen Hendrikse, Paul Nederkoorn, Eline Kooi, Aad van der Lugt, and Marleen de Bruijne. Quantitative comparison of monte-carlo dropout uncertainty measures for multi-class segmentation. In *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Graphs in Biomedical Image Analysis*. 2020.

Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. *arXiv preprint arXiv:2105.05537*, 2021.

Xuyang Cao, Houjin Chen, Yanfeng Li, Yahui Peng, Shu Wang, and Lin Cheng. Uncertainty aware temporal-ensembling model for semi-supervised abus mass segmentation. *IEEE Transactions on Medical Imaging*, 2020.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu. Contrastive learning of global and local features for medical image segmentation with limited annotations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

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.

Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2021a.

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*, 2017.

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European Conference on Computer Vision (ECCV)*, 2018.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, 2020a.

Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021b.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020b.

Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Wenhui Cui, Yanlin Liu, Yuxing Li, Menghao Guo, Yiming Li, Xiuli Li, Tianle Wang, Xiangzhu Zeng, and Chuyang Ye. Semi-supervised brain lesion segmentation with an adapted mean teacher model. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2019a.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019b.

Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Aditya Desai, Zhaozhuo Xu, Menal Gupta, Anu Chandran, Antoine Vial-Aussavy, and Anshumali Shrivastava. Raw nav-merge seismic data to subsurface properties with mlp based multi-modal information unscrambler. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2020.

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*, 2020.

Kang Fang and Wu-Jun Li. Dmnet: difference minimization network for semi-supervised segmentation in medical images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2020.

Gaurav Fotedar, Nima Tajbakhsh, Shilpa Ananth, and Xiaowei Ding. Extreme consistency: Overcoming annotation scarcity and domain shifts. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2020.

Geoff French, Samuli Laine, Timo Aila, Michal Mackiewicz, and Graham Finlayson. Semi-supervised semantic segmentation needs strong, varied perturbations. In *British Machine Vision Conference (BMVC)*, 2020.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning (ICML)*, 2016.

Adrian Galdran, Gustavo Carneiro, and Miguel A González Ballester. Balanced-mixup for highly imbalanced medical image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2021.

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*, 2019.

Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. *Advances in Neural Information Processing Systems (NeurIPS)*, 2004.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006.

Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. *arXiv preprint arXiv:2103.10504*, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Olivier Henaff. Data-efficient image recognition with contrastive predictive coding. In *International Conference on Machine Learning (ICML)*, 2020.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations (ICLR)*, 2019.

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 (MICCAI)*, 2021.

Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 2021.

Ziyu Jiang, Tianlong Chen, Ting Chen, and Zhangyang Wang. Improving contrastive learning on imbalanced data via open-world sampling. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Alain Jungo and Mauricio Reyes. Assessing reliability and challenges of uncertainty estimations for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2019.

Bingyi Kang, Yu Li, Sa Xie, Zehuan Yuan, and Jiashi Feng. Exploring balanced feature spaces for representation learning. In *International Conference on Learning Representations (ICLR)*, 2020.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, 2013.

Junnan Li, Pan Zhou, Caiming Xiong, and Steven Hoi. Prototypical contrastive learning of unsupervised representations. In *International Conference on Learning Representations (ICLR)*, 2021.

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 (MICCAI)*, 2020a.

Xiaomeng Li, Hao Chen, Xiaojuan Qi, Qi Dou, Chi-Wing Fu, and Pheng-Ann Heng. H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes. *IEEE transactions on medical imaging*, 2018.

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*, 2020b.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017a.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision (ICCV)*, 2017b.

Shikun Liu, Shuaifeng Zhi, Edward Johns, and Andrew J Davison. Bootstrapping semantic segmentation with regional contrast. *arXiv preprint arXiv:2104.04465*, 2021.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

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 (MICCAI)*, 2021.

Alireza Mehrtash, William M Wells, Clare M Tempany, Purang Abolmaesumi, and Tina Kapur. Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. *IEEE transactions on medical imaging*, 2020.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *3DV*. IEEE, 2016.

Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Tanya Nair, Doina Precup, Douglas L Arnold, and Tal Arbel. Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. *Medical image analysis*, 2020.

Dong Nie, Yaozong Gao, Li Wang, and Dinggang Shen. Asdnet: Attention based semi-supervised deep networks for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2018.

Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

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

Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Fábio Perez, Cristina Vasconcelos, Sandra Avila, and Eduardo Valle. Data augmentation for skin lesion analysis. In *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis*, pp. 303–311. Springer, 2018.

Siyuan Qiao, Wei Shen, Zhishuai Zhang, Bo Wang, and Alan Yuille. Deep co-training for semi-supervised image recognition. In *European Conference on Computer Vision (ECCV)*, 2018.

Simon Reiß, Constantin Seibold, Alexander Freytag, Erik Rodner, and Rainer Stiefelhagen. Every annotation counts: Multi-label deep supervision for medical image segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

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 (MICCAI)*, 2015.

Abhijit Guha Roy, Jie Ren, Shekoofeh Azizi, Aaron Loh, Vivek Natarajan, Basil Mustafa, Nick Pawlowski, Jan Freyberg, Yuan Liu, Zach Beaver, et al. Does your dermatology classifier know what it doesn't know? detecting the long-tail of unseen conditions. *Medical Image Analysis*, 2022.

Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

Gonglei Shi, Li Xiao, Yang Chen, and S Kevin Zhou. Marginal loss and exclusion loss for partially supervised multi-organ segmentation. *Medical Image Analysis*, 2021.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

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 (NeurIPS)*, 2020.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Ajinkya Tejankar, Soroush Abbasi Koohpayegani, Vipin Pillai, Paolo Favaro, and Hamed Pirsiavash. Isd: Self-supervised learning by iterative similarity distillation. In *IEEE International Conference on Computer Vision (ICCV)*, 2021.

Jeya Maria Jose Valanarasu, Poojan Oza, Ilker Hacihaliloglu, and Vishal M Patel. Medical transformer: Gated axial-attention for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2021.

Gabriele Valvano, Andrea Leo, and Sotirios A Tsaftaris. Learning to segment from scribbles using multi-scale adversarial attention gates. *IEEE Transactions on Medical Imaging*, 2021.

Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, and Luc V Gool. Revisiting contrastive methods for unsupervised learning of visual representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Vikas Verma, Kenji Kawaguchi, Alex Lamb, Juho Kannala, Yoshua Bengio, and David Lopez-Paz. Interpolation consistency training for semi-supervised learning. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2019.

Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Yicheng Wu, Yong Xia, Yang Song, Yanning Zhang, and Weidong Cai. Multiscale network followed network model for retinal vessel segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2018a.

Yicheng Wu, Yong Xia, Yang Song, Donghao Zhang, Dongnan Liu, Chaoyi Zhang, and Weidong Cai. Vessel-net: retinal vessel segmentation under multi-path supervision. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2019.

Zhirong Wu, Yuanjun Xiong, X Yu Stella, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018b.

Jinxi Xiang, Zhuowei Li, Wenji Wang, Qing Xia, and Shaoting Zhang. Self-ensembling contrastive learning for semi-supervised medical image segmentation. *arXiv preprint arXiv:2105.12924*, 2021.

Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020a.

Yutong Xie, Jianpeng Zhang, Zehui Liao, Yong Xia, and Chunhua Shen. Pgl: prior-guided local self-supervised learning for 3d medical image segmentation. *arXiv preprint arXiv:2011.12640*, 2020b.

Yutong Xie, Jianpeng Zhang, Chunhua Shen, and Yong Xia. Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2021.

Zhaozhuo Xu, Zichang Liu, Mauricio Araya-Polo, and Anshumali Shrivastava. Pistachio: Patch importance sampling to accelerate cnns via a hash index optimizer.

Zhaozhuo Xu, Alan Baonan Ji, Andrew Woods, Beidi Chen, and Anshumali Shrivastava. Satellite images and deep learning to identify discrepancy in mailing addresses with applications to census 2020 in houston. *arXiv preprint arXiv:2111.06562*, 2021.

Yuan Xue, Hui Tang, Zhi Qiao, Guanzhong Gong, Yong Yin, Zhen Qian, Chao Huang, Wei Fan, and Xiaolei Huang. Shape-aware organ segmentation by predicting signed distance maps. *arXiv preprint arXiv:1912.03849*, 2019.

Yuzhe Yang and Zhi Xu. Rethinking the value of labels for improving class-imbalanced learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Chenyu You, Ruihan Zhao, Lawrence Staib, and James S Duncan. Momentum contrastive voxel-wise representation learning for semi-supervised volumetric medical image segmentation. *arXiv preprint arXiv:2105.07059*, 2021.

Chenyu You, Ruihan Zhao, Fenglin Liu, Sandeep Chinchali, Ufuk Topcu, Lawrence Staib, and James S Duncan. Class-aware generative adversarial transformers for medical image segmentation. *arXiv preprint arXiv:2201.10737*, 2022a.

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*, 2022b.

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 (MICCAI)*, 2019.

Dewen Zeng, Yawen Wu, Xinrong Hu, Xiaowei Xu, Haiyun Yuan, Meiping Huang, Jian Zhuang, Jingtong Hu, and Yiyu Shi. Positional contrastive learning for volumetric medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2021.

Yu Zeng, Yunzhi Zhuge, Huchuan Lu, and Lihe Zhang. Joint learning of saliency detection and weakly supervised semantic segmentation. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.

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

Zizhao Zhang, Lin Yang, and Yefeng Zheng. Translating and segmenting multimodal medical volumes with cycle-and shape-consistency generative adversarial network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Han Zheng, Lanfen Lin, Hongjie Hu, Qiaowei Zhang, Qingqing Chen, Yutaro Iwamoto, Xianhua Han, Yen-Wei Chen, Ruofeng Tong, and Jian Wu. Semi-supervised segmentation of liver using adversarial learning with deep atlas prior. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2019.

Mingkai Zheng, Shan You, Fei Wang, Chen Qian, Changshui Zhang, Xiaogang Wang, and Chang Xu. Ressl: Relational self-supervised learning with weak augmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Yanning Zhou, Hao Chen, Huangjing Lin, and Pheng-Ann Heng. Deep semi-supervised knowledge distillation for overlapping cervical cell instance segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2020.

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

Xiangxin Zhu, Dragomir Anguelov, and Deva Ramanan. Capturing long-tail distributions of object subcategories. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

Xiahai Zhuang and Juan Shen. Multi-scale patch and multi-modality atlases for whole heart segmentation of mri. *Medical image analysis*, 2016.

Appendix to
Mine yOur owN Anatomy: Revisiting Medical Image Segmentation with Extremely Limited Labels

section A provides additional training details.

section B provides more experimental results on MMWHS.

section C compares to existing state-of-the-art contrastive learning (CL) frameworks.

section D provides more ablations on anatomical contrastive reconstruction.

A More Training Details

The ACDC dataset was hosted in MICCAI 2017 ACDC challenge Bernard et al. (2018), which includes 200 3D cardiac cine MRI scans with expert annotations for three classes (*i.e.*, left ventricle (LV), myocardium (Myo), and right ventricle (RV)). We divide the dataset into splits of 120, 40 and 40 scans for training, validation, and testing with a random order. For pre-processing, we adopt the similar setting in Chaitanya et al. (2020) by normalizing the intensity of each 3D scan (*i.e.*, using min-max normalization) into $[0, 1]$, and re-sampling all 2D scans and the corresponding segmentation maps into a fixed spatial resolution of 256×256 pixels.

The LiTS dataset was hosted in MICCAI 2017 Liver Tumor Segmentation Challenge Bilic et al. (2019), which includes 131 contrast-enhanced 3D abdominal CT volumes with expert annotations for two classes (*i.e.*, liver and tumor). We divide the dataset into splits of 100 and 31 scans for training and testing with a random order. For pre-processing, we adopt the similar setting in Li et al. (2018) by truncating the intensity of each 3D scan into $[-200, 250]$ HU for removing irrelevant and redundant details, normalizing each 3D scan into $[0, 1]$, and re-sampling all 2D scans and the corresponding segmentation maps into a fixed spatial resolution of 256×256 pixels.

The MMWHS dataset was hosted in MICCAI 2017 challenge Zhuang & Shen (2016), which includes 20 3D cardiac MRI scans with expert annotations for seven classes: left ventricle (LV), left atrium (LA), right ventricle (RV), right atrium (RA), myocardium (Myo), ascending aorta (AAo), and pulmonary artery (PA). We divide the dataset into splits of 15 and 5 scans for training and testing with a random order. For pre-processing, we normalize the intensity of each 3D scan (*i.e.*, using min-max normalization) into $[0, 1]$, and re-sampling all 2D scans and the corresponding segmentation maps into a fixed spatial resolution of 256×256 pixels.

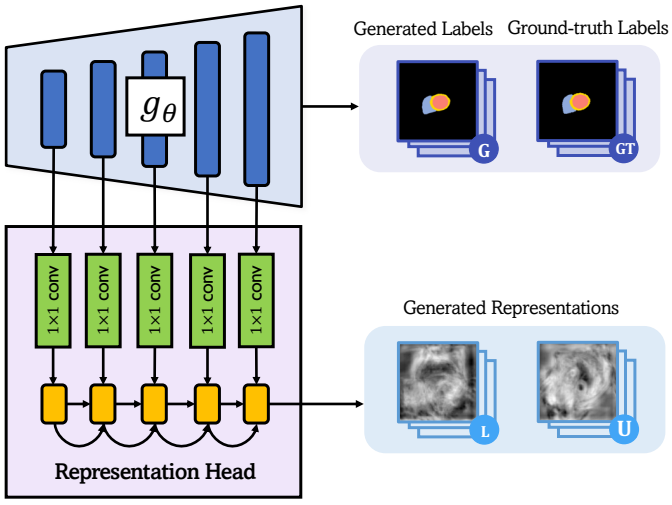


Figure 5: Representation head architecture.

Implementation details. We implement all the evaluated models using PyTorch library Paszke et al. (2019). All the models are trained using Stochastic Gradient Descent (SGD) (*i.e.*, initial learning rate

= 0.01, momentum = 0.9, weight decay = 0.0001) with batch size of 6, and the initial learning rate is divided by 10 every 2500 iterations. All of our experiments are conducted on NVIDIA GeForce RTX 3090 GPUs. We first train our model with 100 epochs during the pre-training, and then retrain the model for 300 epochs during the fine-tuning. We set the temperature τ_ξ, τ_θ, τ as 0.01, 0.1, 0.5. The size of the memory bank is 36. During the pre-training, we follow the settings of `ISD`, including global projection head setting, and predictors with the 512-dimensional output embedding, and adopt the setting of local projection head in Hu et al. (2021). More specifically, given the predicted logits $\hat{\mathbf{y}} \in \mathbb{R}^{\mathcal{C}\times\mathcal{H}\times\mathcal{W}}$, we create 36 different views (*i.e.*, random crops at the same location) of $\hat{\mathbf{y}}$ and $\hat{\mathbf{y}}'$ with the fixed size 64×64, and then project all pixels into 512-dimensional output embedding space, and the output feature dimension of h'_θ is also 512. An illustration of our representation head is presented in Figure 5. We then actively sample 256 query embeddings and 512 key embeddings for each mini-batch, and the confidence threshold δ_θ is set to 0.97. When fine-tuning we use an equally sized pool of candidates $K = 5$, as well as $\lambda_1 = 0.01$, $\lambda_2 = 1.0$, $\lambda_3 = 1.0$, and $\lambda_4 = 1.0$. For different augmentation strategies, we implement the weak augmentation to the student's input as random rotation, random cropping, horizontal flipping, and strong augmentation to the teacher's input as random rotation, random cropping, horizontal flipping, random contrast, CutMix French et al. (2020), brightness changes Perez et al. (2018), morphological changes (diffeomorphic deformations). We adopt two popular evaluation metrics: Dice coefficient (DSC) and Average Symmetric Surface Distance (ASD) for 3D segmentation results. Of note, the projection heads, the predictor, and the representation head are only used in training, and will be discarded during inference.

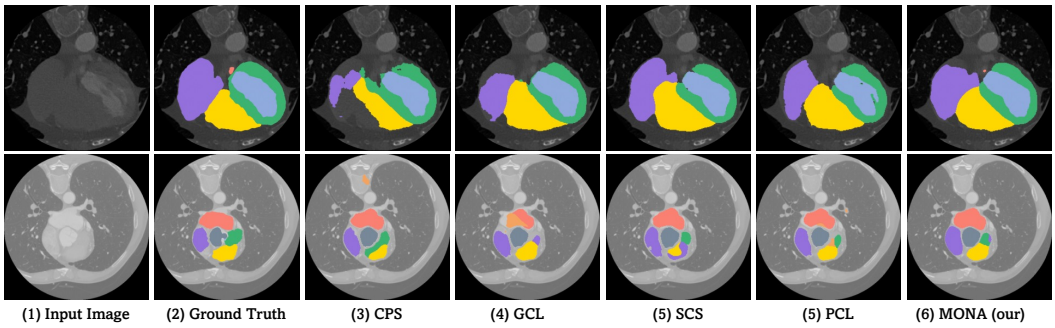


Figure 6: Visualization of segmentation results on MMWHS with 5% labeled ratio. As is shown, `MONA` consistently generates more accurate predictions compared to all other SSL methods with a significant performance margin. Different anatomical classes are shown in different colors (LV: ; LA: ; RV: ; RA: ; Myo: ; PA:).

B More Experiments Results - MMWHS

Lastly, we validate `MONA` on MMWHS, under 1%, 5%, 10% labeled ratios. The results are provided in Table 4 and Figure 6. Again, we found that `MONA` consistently outperforms all other SSL methods with a significant performance margin, and achieves the highest accuracy among all the SSL and fully supervised approaches under three labeled ratios. As is shown, `MONA` trained at the 1% labeled ratio significantly outperforms all other methods trained at the 1% labeled ratio, even over the 5% labeled ratio. Concretely, `MONA` trained at only 1% labeled ratio outperforms the second-best method (*i.e.*, `GLCon`) both at the 1% and 5% labeled, yielding 12.3% and 0.4% gains in Dice. We also observe the similar patterns that, `MONA` performs better or on par with all the other methods at 10% labeled. Particularly, `MONA` trained at both 5% and 10% labeled ratio surpasses the fully-supervised scheme by 0.6% and 1.8% improvements in Dice, which again demonstrates the superiority of `MONA` in extremely limited labeled data regimes.

C Generalization Study of Contrastive Learning Pre-training

As discussed in Section 3.1, our motivation comes from the observation that there are only very limited labeled data and a large amount of unlabeled data in real-world clinical practice. As the fully-supervised methods generally outperform all other SSL methods by clear margins, we postulate

Table 4: Comparison of segmentation performance (DSC[%]/ASD[mm]) on MMWHS under three labeled ratio settings (1%, 5%, 10%). On all three labeled settings, MONA significantly outperforms all the state-of-the-art methods by a significant margin. The best results are in **bold**.

Method	1% Labeled		5% Labeled		10% Labeled	
	DSC ↑	ASD ↓	DSC ↑	ASD ↓	DSC ↑	ASD ↓
UNet-F Ronneberger et al. (2015)	85.8	8.01	85.8	8.01	85.8	8.01
UNet-L	58.3	33.9	77.8	24.4	82.7	13.5
EM Vu et al. (2019)	54.5	41.1	80.6	17.3	82.1	15.1
CCT Ouali et al. (2020)	62.8	27.5	79.0	21.9	79.4	16.3
DAN Zhang et al. (2017)	52.8	48.4	79.4	22.7	80.2	15.0
URPC Luo et al. (2021)	65.7	29.7	73.7	20.5	81.9	12.3
DCT Qiao et al. (2018)	62.7	27.5	80.8	23.0	82.8	12.4
ICT Verma et al. (2019)	59.9	32.8	76.5	15.4	82.2	12.0
MT Tarvainen & Valpola (2017)	58.8	35.6	76.5	15.5	79.4	19.8
UAMT Yu et al. (2019)	61.1	37.6	76.3	20.9	83.7	14.2
CPS Chen et al. (2021b)	58.8	33.6	78.3	22.5	82.0	13.1
GCL Chaitanya et al. (2020)	71.6	20.3	83.5	**7.41**	86.7	8.76
SCS Hu et al. (2021)	71.4	19.3	81.1	11.5	82.6	9.68
PLC Chaitanya et al. (2021)	71.5	19.8	83.4	10.7	86.0	9.65
• MONA (ours)	**83.9**	**9.06**	**86.3**	8.22	**87.6**	**6.83**

that leveraging massive unlabeled data usually introduces additional noise during training, leading to degraded segmentation quality. To address this challenge, "contrastive learning" is a straightforward way to leverage existing unlabeled data in the learning procedure. As supported in Section 4 and Appendix B, our findings have shown that MONA generalizes well across different benchmark datasets (*i.e.*, ACDC, LiTS, MMWHS) with diverse labeled settings (*i.e.*, 1%, 5%, 10%). In the following subsection, we further demonstrate that our proposed principles (*i.e.*, tailness, consistency, diversity) are beneficial to various state-of-the-art CL-based frameworks (*i.e.*, MoCov2 Chen et al. (2020b), kNN-MoCo Van Gansbeke et al. (2021), SimCLR Chen et al. (2020a), BYOL Grill et al. (2020), and ISD Tejankar et al. (2021)) with different label settings. More details about these three principles can be found in Section 3.2. Of note, to the best of our knowledge, MONA is the first SSL training scheme that consistently outperforms the fully-supervised method on diverse benchmark datasets with only 10% labeled ratio.

Training details of competing CL methods. We identically follow the default setting in each CL framework Chen et al. (2020b); Van Gansbeke et al. (2021); Chen et al. (2020a); Grill et al. (2020); Tejankar et al. (2021) except the epochs number. We train each model in the semi-supervised setting. For labeled data, we follow the same training strategy in Section 3.1. As for unlabeled data, we strictly follow the default settings in each baseline. Specifically, for fair comparisons, we pre-train each CL baseline and our proposed CL pre-trained method (*i.e.*, GLCon) for 100 epochs in all our experiments. Then we fine-tune each CL model with our proposed principles with the same setting, as provided in Appendix A. For kNN-MoCo Van Gansbeke et al. (2021), given the following ablation study we set the number of neighbors k as 5, and further compare different settings of k in kNN-MoCo Van Gansbeke et al. (2021) in the following subsection. All the experiments are run with three different random seeds, and the results we present are calculated from the validation set.

Comparisons with CL-based frameworks. Table 5 presents the comparisons between our proposed methods (*i.e.*, GLCon and MONA) and various CL baselines. After analyzing these extensive results, we can draw several consistent observations. ***First***, we can observe that our proposed GLCon achieves performance gains under all the labeled ratios, which not only demonstrates the effectiveness of our method, but also further verifies this argument using "global-local" strategy Chaitanya et al. (2020). The average improvement in Dice obtained by GLCon could reach up to 2.53%, compared to the second best scores at different labeled ratios. ***Second***, we can find that incorporating our proposed three principles significantly outperforms the CL baselines without fine-tuning, across all frameworks and different labeled ratios. These experimental findings suggest that our proposed three principles can further improve the generalization across different labeled ratios. On the ACDC dataset at the 1% labeled ratio, the backbones equipped with all three principles all obtain promising results, improving the performance of MoCov2, kNN-MoCo, SimCLR, BYOL, ISD, and our GLCon by 39.1%, 38.5%, 40.9%, 41.2%, 34.3%, 34.0%, respectively. The ACDC dataset is a popular multi-class medical image segmentation dataset, with massive imbalanced or long-tailed class distribution cases. The imbalanced or long-tailed class distribution gap could result in the vanilla models overfitting to the head class, and generalizing very poorly to the tail class. With the addition of under-sampling

Table 5: Ablation study of different contrastive learning frameworks on ACDC under three labeled ratio settings (1%, 5%, 10%). We compare two settings: with or without *fine-tuning* on the segmentation performance (DSC[%]/ASD[mm]). We denote 'without *fine-tuning*" to only *pretaining*. On all three labeled settings, our methods (*i.e.*, GLCon and MONA) significantly outperform all the state-of-the-art methods by a significant margin. All the experiments are run with three different random seeds. The best results are in **bold**.

Framework	Method	1% Labeled		5% Labeled		10% Labeled	
		DSC ↑	ASD ↓	DSC ↑	ASD ↓	DSC ↑	ASD ↓
only *pretaining*	MoCov2 Chen et al. (2020b)	38.6	22.4	56.2	17.9	81.0	5.36
	kNN-MoCo Van Gansbeke et al. (2021)	39.5	22.0	58.3	15.7	83.1	7.18
	SimCLR Chen et al. (2020a)	34.8	24.3	51.7	19.9	80.3	4.16
	BYOL Grill et al. (2020)	35.9	7.25	65.9	9.15	85.6	2.51
	ISD Tejankar et al. (2021)	45.8	17.2	71.0	4.29	85.3	2.97
	• GLCon (ours)	**49.3**	**7.11**	**74.2**	**3.89**	**86.5**	**1.92**
w/ *fine-tuning*	MoCov2 Chen et al. (2020b)	77.7	4.78	85.4	1.52	86.7	1.74
	kNN-MoCo Van Gansbeke et al. (2021)	78.0	4.28	85.9	1.51	86.9	1.61
	SimCLR Chen et al. (2020a)	75.7	4.33	83.2	2.06	86.1	2.25
	BYOL Grill et al. (2020)	77.1	4.84	85.3	2.06	88.1	0.99
	ISD Tejankar et al. (2021)	80.1	3.00	83.8	1.95	88.6	1.20
	• MONA (ours)	**83.3**	**1.98**	**89.1**	**0.784**	**90.8**	**0.736**

the head classes, the principle – *tailness* – can be deemed as the prominent strategy to yield better generalization and segmentation performance of the models across different labeled ratios. Similar results are found under 5% and 10% labeled ratios. *Third*, over a wide range of labeled ratios, MONA can establish the new state-of-the-art performance bar for semi-supervised medical image segmentation. Particularly, MONA – for the first time – boots the segmentation performance with 10% labeled ratio over the fully-supervised method while significantly outperforming all the other semi-supervised methods by a large margin. In summary, our proposed methods (*i.e.*, GLCon and MONA) obtain remarkable performance on all labeled settings. The results verify the superiority of our proposed three principles (*i.e.*, tailness, consistency, diversity) jointly, which makes the model well generalize to different labeled settings, and can be easily and seamlessly plugged into all other CL frameworks Chen et al. (2020b); Van Gansbeke et al. (2021); Chen et al. (2020a); Grill et al. (2020); Tejankar et al. (2021) adopting the two-branch design, demonstrating that these concepts consistently help the model yield extra performance boosts for them all.

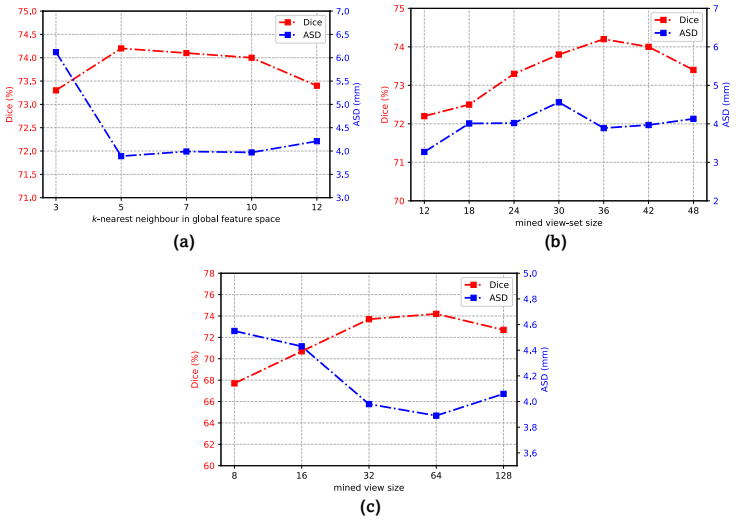


Figure 7: Effects of k-nearest neighbour in global feature space, mined view-set size, and mined view patch size. We report Dice and ASD of GLCon on the ACDC dataset at the 5% labeled ratio. All the experiments are run with three different random seeds.

Does k-nearest neighbour in global feature space help? Prior work suggests that the use of stronger augmentations and nearest neighbour can be the very effective tools in learning additional invariances Van Gansbeke et al. (2021). That is, both the specific number of nearest neighbours and specific

augmentation strategies are necessary to achieve superior performance. In this subsection, we study the relationship of k-nearest neighbour in global feature space and the behavior of our GLCon for the downstream medical image segmentation. Here we first follow the same augmentation strategies in Van Gansbeke et al. (2021) (More analysis on data augmentation can be found in Section 4.2), and then conduct ablation studies on how the choices of k-nearest neighbour can influence the performance of GLCon. Specifically, we run GLCon on the ACDC dataset at the 5% labeled ratio with a range of $k \in \{3, 5, 7, 10, 12\}$. Figure 7(a) shows the ablation study on k-nearest neighbour in global feature on the segmentation performance. As is shown, we find that GLCon at $k = 5, 7, 10$ have almost identical performance ($k = 5$ has slightly better performance compared to other two settings), and all have superior performance compared to all others. In contrast, GLCon – through the use of randomly selected samples – is capable of finding diverse yet semantically consistent anatomical features from the entire dataset, which at the same time gives better segmentation performance.

Ablation study of mined view-set size. We then conduct ablation studies on how the mined view-set size in GLCon can influence the segmentation performance. We run GLCon on the ACDC dataset at 5% labeled ratio with a range of the mined view-set size $\in \{12, 18, 24, 30, 36, 42, 48\}$. The results are summarized in Figure 7(b). As is shown, we find that GLCon trained with view-set size 36 and 42 have similar or superior performance compared to all other settings, and our model with view-set size of 36 achieves the highest performance.

Ablation study of mined view size. Lastly, we study the influence of mined view size on the segmentation performance. Specifically, we run GLCon on the ACDC dataset at the 5% labeled ratio with a range of the mined view size $\in \{8, 16, 32, 64, 128\}$. Figure 7(c) shows the ablation study of mined view size on the segmentation performance. As is shown, we observe that GLCon trained with mined view size of 32 and 64 have similar segmentation abilities, and both achieve superior performance compared to other settings. Here the mined view size of 64 works the best for GLCon to yield the superior segmentation performance.

Conclusion. Given the above ablation study, we set k, mined view-set size, patch size as 5, 36, 64×64 in our experiments, respectively. This can contribute to satisfactory segmentation performance.

D Ablation Study of Anatomical Contrastive Reconstruction

In this section, we give a detailed analysis on the choice of the parameters in the anatomical contrastive reconstruction fine-tuning, and take a deeper look and understand how they contribute to the final segmentation performance. All the hyperparameters in training are the same across three benchmark datasets. All the experiments are run with three different random seeds, and the experimental results we report are calculated from the validation set.

Ablation study of total loss $\mathcal{L}_{\text{total}}$. Proper choices of hyperparameters in total loss $\mathcal{L}_{\text{total}}$ (See Section 3.2) play a significant role in improving overall segmentation quality. We hence conduct the fine-grained analysis of the hyperparameters in $\mathcal{L}_{\text{total}}$. In practice, we fine-tune the models with three independent runs, and grid search to select multiple hyperparameters. Specifically, we run MONA on the ACDC dataset at the 5% labeled ratio with a range of different hyperparameters $\lambda_1 \in \{0.005, 0.001, 0.05, 0.01, 0.05, 0.1\}$, and $\lambda_2, \lambda_3, \lambda_4 \in \{0.1, 0.2, 0.5, 1.0, 2.0, 10.0\}$. We summarize the results in Figure 8, and take the best setting $\lambda_1 = 0.01$, $\lambda_2 = 1.0$, $\lambda_3 = 1.0$, $\lambda_4 = 1.0$.

Ablation study of confidence threshold δ_θ. We then assess the influence of δ_θ on the segmentation performance. Specifically, we run MONA on the ACDC dataset at the 5% labeled ratio with a range of the confidence threshold $\delta_\theta \in \{0.85, 0.88, 0.91, 0.94, 0.97, 1.0\}$. Figure 9(a) shows the ablation study of δ_θ on the segmentation performance. As we can see, MONA on $\delta_\theta = 0.97$ has superior performance compared to other settings.

Ablation study of K-nearest neighbour constraint. Next, we conduct ablation studies on how the choices of K in K-nearest neighbour constraint can influence the segmentation performance. Specifically, we run MONA on the ACDC dataset at the 5% labeled ratio with a range of the choices $K \in \{3, 5, 7, 10, 12\}$. Figure 9(b) shows the ablation study of K choices on the segmentation performance. As we can see, MONA on $K = 5$ achieves the best performance compared to other settings.

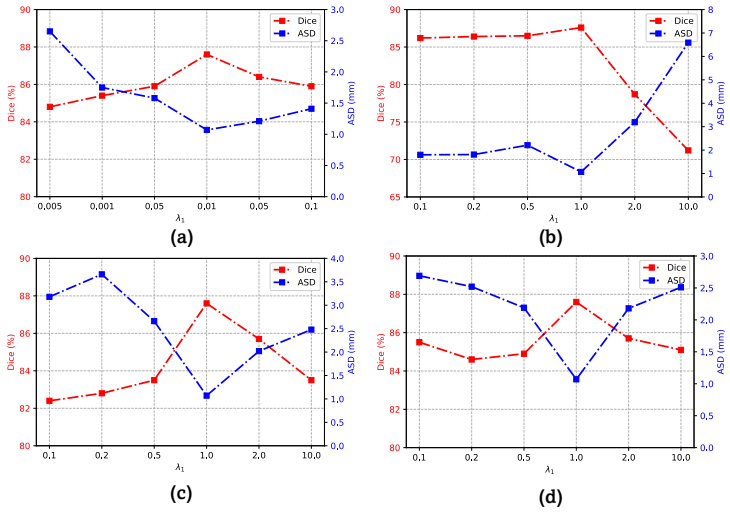


Figure 8: Effects of hyperparameters $\lambda_1, \lambda_2, \lambda_3, \lambda_4$. We report Dice and ASD of MONA on the ACDC dataset at the 5% labeled ratio. All the experiments are run with three different random seeds.

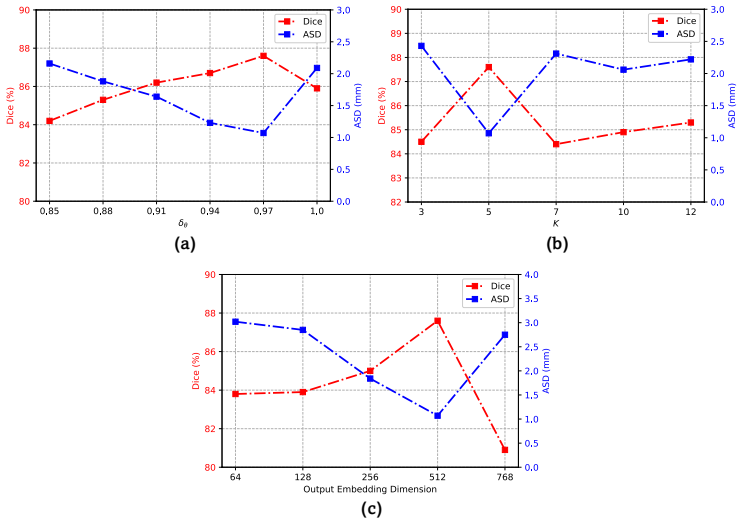


Figure 9: Effects of confidence threshold δ_θ, K-nearest neighbour constraint, and output embedding dimension. We report Dice and ASD of MONA on the ACDC dataset at the 5% labeled ratio. All the experiments are run with three different random seeds.

Ablation study of output embedding dimension. Finally, we study the influence of the output embedding dimension on the segmentation performance of MONA. Specifically, we run MONA on the ACDC dataset at the 5% labeled ratio with a range of output embedding dimension $\in \{64, 128, 256, 512, 768\}$. Figure 9(c) shows the ablation study of output embedding dimension on the segmentation performance. As we can see, MONA with output embedding dimension of 512, can be trained to outperform other settings.

Conclusion. Given the above ablation study, we select $\lambda_1 = 0.01$, $\lambda_2 = 1.0$, $\lambda_3 = 1.0$, $\lambda_4 = 1.0$, $\delta_\theta = 0.97$, $K = 5$, output embedding dimension $= 512$ in our experiments. This can provide the optimal segmentation performance across different labeled ratios.