Mine yOur owN Anatomy: Revisiting Medical Image Segmentation With Extremely Limited Labels

Chenyu You , Weicheng Dai, Fenglin Liu , Yifei Min, Nicha C. Dvornek , Xiaoxiao Li , David A. Clifton , Lawrence Staib , and James S. Duncan , *Life Fellow, IEEE*

Abstract—Recent studies on contrastive learning have achieved remarkable performance solely by leveraging few labels in medical image segmentation. Existing methods mainly focus on instance discrimination and invariant mapping. 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 Mine yOur owN Anatomy (MONA), and make three contributions. First, prior work argues that every pixel equally matters to the 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. 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, we both empirically and theoretically, demonstrate the efficacy of our MONA on three benchmark datasets with different labeled settings, achieving new state-of-the-art under different labeled semi-supervised settings.

Index Terms—Contrastive learning, imbalanced learning, long-tailed medical image segmentation, semi-supervised learning.

Chenyu You, Weicheng Dai, Yifei Min, Lawrence Staib, and James S. Duncan are with the Yale University, New Haven, CT 06510 USA (e-mail: chenyu.you@yale.edu; weicheng.dai@yale.edu; yifei.min@yale.edu; lawrence.staib@yale.edu; james.duncan@yale.edu).

Fenglin Liu is with the University of Oxford, OX3 7DQ Oxford, U.K. (e-mail: fenglin.liu@eng.ox.ac.uk).

Nicha C. Dvornek is with Yale University, New Haven, CT 06510 USA (e-mail: nicha.chitphakdithai@yale.edu).

Xiaoxiao Li is with the University of British Columbia, Vancouver, BC V6T 1Z4, Canada (e-mail: xiaoxiao.li@ece.ubc.ca).

David A. Clifton is with the University of Oxford, OX3 7DQ Oxford, U.K., and also with the Oxford-Suzhou Centre for Advanced Research, Suzhou 215000, China (e-mail: david.clifton@eng.ox.ac.uk).

Codes will be available on here. https://github.com/charlesyou999648/MONA

I. 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 [1], [2], [3], [4], [5], [6], [7], [8], [9], [10]. 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.

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 [11], [12], [13]. 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 [14], [15], [16], [17], [18]. In contrast, there have been few prior long-tail works for medical image segmentation. For example, as illustrated in Fig. 1, most medical 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 [19], [20], [21] pointed out that going beyond simple augmentations to create more diverse views can learn more discriminative anatomical features. At
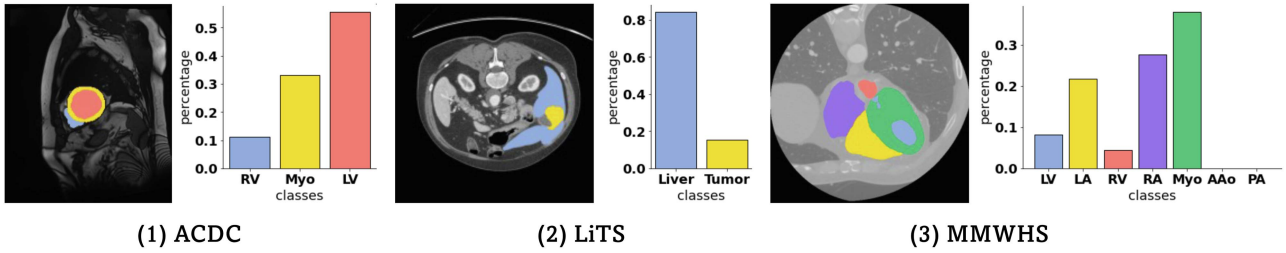
Fig. 1. Examples of three benchmarks (i.e., ACDC, LiTS, MMWHS) with long-tail class distributions. As observed, the ratios of different label classes over three benchmarks are imbalanced.
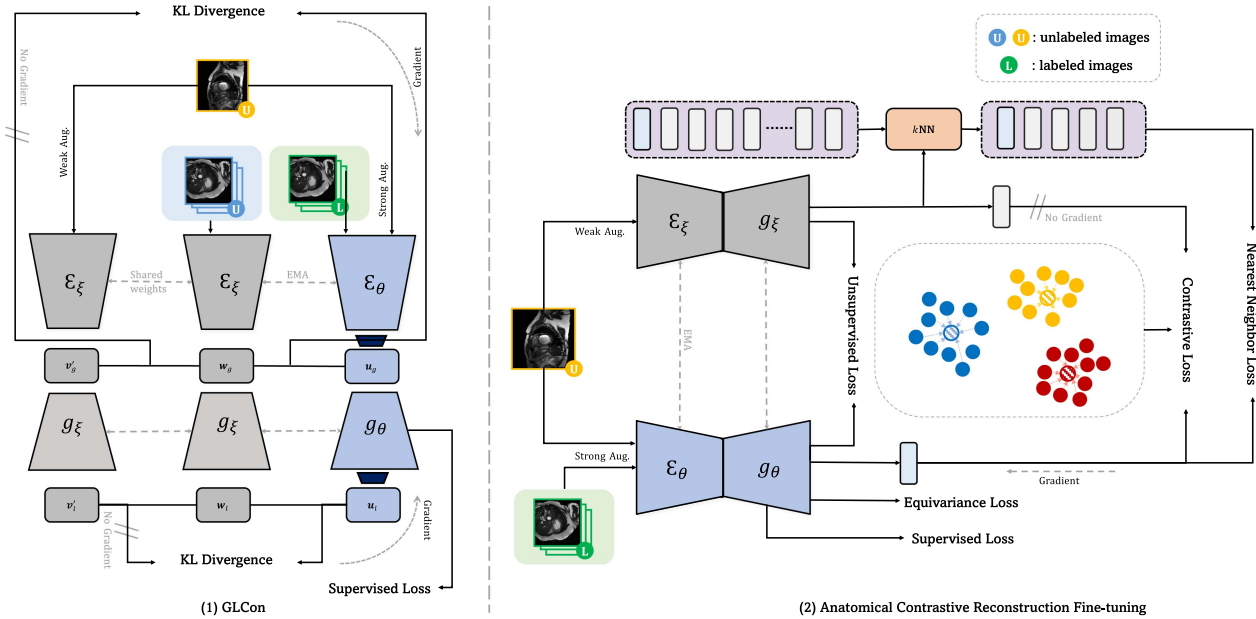

Fig. 2. Overview of the MONA framework including two stages: (1) GLCon is design to seek both *augmented* and *mined* views for instance discrimination $\mathcal{L}_{inst}$ in the global and local manners. Here the global instance discrimination is designed to exploit the correlations among views within the latent feature space, which is generated by the encoders. Meanwhile, local instance discrimination aims to leverage the correlations among views - specifically, local regions of the image - within the output feature space produced by the decoder (See Section III-A), (2) our proposed anatomical contrastive reconstruction fine-tuning (See Section III-B). Note that U and L denote unlabeled and labeled data.

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 2D medical segmentation framework, based on different anatomical views. The workflow of MONA is illustrated in Fig. 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 [4], [5], [6] 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 [22] 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 [23] and instance discrimination [1], [4]. 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 [5], SimCLR [4], BYOL [6], and ISD [24]), which consistently demonstrate the effectiveness of our proposed MONA. For example, our MONA establishes the **new state-of-the-art** performance, compared to all the state-of-the-art semi-supervised approaches with different label ratios (i.e., 1%, 5%, 10%). Moreover, we present a systematic evaluation for analyzing why our approach performs so well and how different factors contribute to the final performance (See Section IV-D). Theoretically, we show the efficacy of our MONA in label efficiency (See Section A). Empirically, we also study whether these principles can effectively complement each CL framework (See Section IV-G). We hope our findings will provide useful insights on medical image segmentation to other researchers.

To summarise, our contributions are as follows: ❶ we carefully examine the problem of semi-supervised 2D medical image segmentation with extremely limited labels, and identify the three principles to address such challenging tasks; ❷ we construct a set of objectives, which significantly improves the segmentation quality, both long-tail class distribution and anatomical features; ❸ we both empirically and theoretically analyze several critical components of our method and conduct thorough ablation studies to validate their necessity; ❹ with the combination of different components, we establish state-of-the-art under SSL settings, for all the challenging three benchmarks.

II. 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) [25] and UNET [26]. Follow-up works can be categorized into two main directions. One direction is to improve modern segmentation network design. Many CNN-based [27], [28] and Transformer-like [29], [30] model variants [31], [32], [33], [34], [35], [36], [37], [38], [39], [40], [41] have been proposed since then. For example, some works [32], [35], [42] proposed to use dilated/atrous/deformable convolutions with larger receptive fields for more dense anatomical features. Other works [36], [37], [38], [39], [40], [41] 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 [43], [44], [45]. 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 [46], [47], [48], [49], [50], model uncertainty [51], [52], [53], confidence estimation [54], [55], [56], and noisy student [57]; (2) consistency regularization [58], [59], [60] by integrating consistency corresponding to different transformation, such as pi-model [61], co-training [62], [63], and mean-teacher [9], [10], [64], [65], [66], [67]; (3) other training strategies such as adversarial training [68], [69], [70], [71], [72], [73] and entropy minimization [74]. 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 [9], [10], [75], [76], [77], [78], [79], [80], [81], [82]. The high-level idea of CL is to pull closer the different 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*. [83] augments a given image with 4 different rotation degrees and trains the model to be aware of which rotation degree of each image by applying an contrastive loss. In contrast, our goal is to train a model to yield segments that adhere to anatomical, geometric and equivariance constraints in an unsupervised manner. A few very recent studies [14], [18] confirm the superiority of CL of addressing imbalance issues in image classification. Moreover, existing CL frameworks [75], [77] 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 [84], [85], [86]. Inspired by the aforementioned, we address these two issues in medical image segmentation - two appealing perspectives that still remain under-explored.

III. MINE YOUR OWN ANATOMY (MONA)

Overview: MONA consists of two parts: a global-local contrastive pre-training part named GLCon (Section III-A) and a fine-tuning part named Anatomical Contrastive Reconstruction (Section III-B). We illustrate our contrastive learning framework (See Fig. 2), which includes (1) relational semi-supervised

pre-training, and (2) anatomical contrastive reconstruction fine-tuning.

A. GLCon

Our pre-training stage is built upon ISD [24] - a competitive framework for image classification. The *main differences* between ISD and the pre-training part of MONA (i.e. GLCon) are: GLCon 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 [5], SimCLR [4], BYOL [6]). More detailed empirical and theoretical analysis can be found in Section IV-G and Section A, available online.

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 [26]). 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 [75], we adopt the global and local instance discrimination strategies in the latent and output feature spaces.[2] Specifically, the encoders generate global features $\mathbf{z}_g = \mathcal{E}_\theta(\mathbf{x})$, $\mathbf{z}'_g = \mathcal{E}_\xi(\mathbf{x}')$, and $\mathbf{z}_{r,g} = \mathcal{E}_\xi(\mathbf{x}_r)$, which are then fed into the nonlinear projection heads to obtain $\mathbf{v}_g = h_\theta(\mathbf{z}_g)$, $\mathbf{v}'_g = h_\xi(\mathbf{z}'_g)$, 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., $\mathbf{u}_g = h'_\theta(\mathbf{v}_g)$. 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:

$$\mathbf{s}_\theta = \log \frac{\exp\left(\text{sim}\left(\mathbf{u}, \mathbf{w}\right)/\tau_\theta\right)}{\sum_{j=1}^{k} \exp\left(\text{sim}\left(\mathbf{u}, \mathbf{w}_j\right)/\tau_\theta\right)},$$

$$\mathbf{s}_\xi = \log \frac{\exp\left(\text{sim}\left(\mathbf{v}', \mathbf{w}\right)/\tau_\xi\right)}{\sum_{j=1}^{k} \exp\left(\text{sim}\left(\mathbf{v}', \mathbf{w}_j\right)/\tau_\xi\right)}, \quad (1)$$

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

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

---

[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.
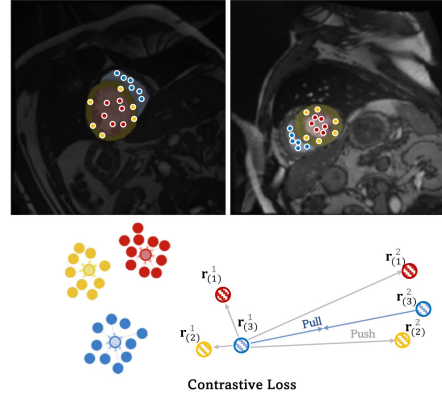

Fig. 3. Illustration of the contrastive loss. Intuitively, we actively sample a set of pixel-level anchor representations, pulling them closer to the class-averaged mean of representations within this class (*positive keys*), and pushing away from representations from other classes (*negative keys*).

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): $\mathcal{L}_{\text{inst}}^{\text{global}} + \mathcal{L}_{\text{inst}}^{\text{local}} + \mathcal{L}_{\text{sup}}$. Therefore, the GLCon loss encourages that the model acquires both global and local features.

B. Anatomical Contrastive Reconstruction

Principles: The key idea of the fine-tuning part is to seek diverse yet semantically consistent views whose anatomical features are *homogeneous* within the *same class type*, while *distinctive* for *different class types*. As shown in Fig. 2, the principles behind MONA (the 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. More theoretical analysis is in Section A, available online.

Tailness: Motivated by the observations (Fig. 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**) (See Fig. 3).

Here we additionally attach the representation heads to fuse the multi-scale features with the feature pyramid network (FPN) [87] 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 $\mathbf{r}_q \in \mathcal{R}_q^c$ (*queries*), pulling them closer to the class-averaged mean of representations $\mathbf{r}_k^{c,+}$ within this class c (*positive keys*), and pushing away from representations $\mathbf{r}_k^- \in \mathcal{R}_k^c$ from other classes (*negative keys*). Formally, the contrastive loss is defined as:

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

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:

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

Note that in (3), we are using the negative pairs $\mathbf{r}_k^-$ to estimate the centers of opposite classes. The class average representation $\mathbf{r}_k^{c,+}$ is averaged over all instances from the target class c. We also 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. In order to compare the pairs of instances between opposite and target classes, we then create a graph $\mathcal{G}$ to compute the pair-wise class relationship: $\mathcal{G}[p,q] = (\mathbf{r}_k^{p,+} \cdot \mathbf{r}_k^{q,+}), \forall p, q \in \mathcal{C}$, 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 be specific, in (3), we use adaptive sampling for the negative keys $\mathbf{r}_k^-$ from the opposite classes. To do so, we use `softmax` to yield a distribution $\exp(G[c,v])/\sum_{n \in \mathcal{C}, n \neq c} \exp(G[c,n])$, with which we adaptively sample negative keys from class v, for $v \neq c$. To address the challenge in *imbalanced* medical image data, we define the pseudo-label (i.e., easy and hard queries) based on a
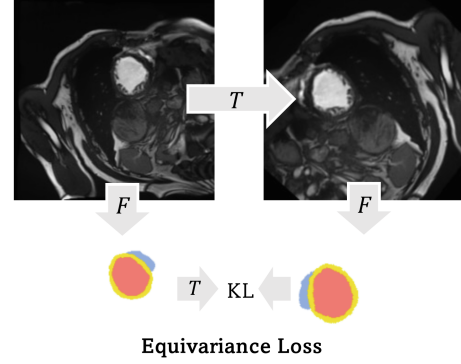

Equivariance Loss

Fig. 4. Illustration of the equivariance loss.

defined threshold as follows:

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

where $\hat{\mathbf{y}}_q$ is the cth-class pseudo-label corresponding to $\mathbf{r}_q$, and δ_θ is the user-defined threshold. For further improvement in long-tail scenarios, we construct a class-aware memory bank [5] 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. Our goal is to train a model to yield segments that adhere to anatomical, geometric and equivariance constraints in an unsupervised manner. As shown in Fig. 4, 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:

$$
\mathcal{L}_{\text{eqv}}(\mathbf{x}, \mathcal{T}(\mathbf{x})) = \sum_{\mathbf{x} \in X} \mathcal{KL}(\mathcal{T}(\mathcal{F}_\theta(\mathbf{x})), \mathcal{F}_\theta(\mathcal{T}(\mathbf{x})))
$$
$$
+ \mathcal{KL}(\mathcal{F}_\theta(\mathcal{T}(\mathbf{x})), \mathcal{T}(\mathcal{F}_\theta(\mathbf{x}))).
$$
(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 Section IV-H). To counteract this, we utilize the nearest neighbor strategy, ensuring the model benefits from its previous outputs without overly concentrating on extraneous features. Thus, we formulate our insight as an auxiliary loss that regularizes the representations

- keeping the anatomical contrastive reconstruction task as the main force. In practice, given a batch of unlabeled images, we use both the teacher and student models to obtain $\mathbf{v}'_g$ and $\mathbf{u}_g$, which are then normalized using the l_2 norm. $\mathbf{v}'_g$ is fed to the first-in-first-out (FIFO) memory bank [5], where it search for K-nearest neighbors from the memory bank. Then we use the nearest neighbor loss $\mathcal{L}_{nn}$ to maximize cosine similarity, thereby exploiting the inter-instance relationship. Specifically, we minimize the distance between $\mathbf{u}_g$ and the K-nearest neighbors, with the distance defined as negative cosine similarity, thereby maximizing cosine similarity.

Setup: The total loss $\mathcal{L}_{total}$ is the sum of contrastive loss $\mathcal{L}_{contrast}$ (on both ground-truth labels and pseudo-labels), equivariance loss $\mathcal{L}_{eqv}$ (on both ground-truth labels and pseudo-labels), nearest neighbors loss $\mathcal{L}_{nn}$ (on both ground-truth labels and pseudo-labels), unsupervised cross-entropy loss $\mathcal{L}_{unsup}$ (on pseudo-labels) and supervised segmentation loss $\mathcal{L}_{sup}$ (on ground-truth labels): $\mathcal{L}_{sup} + \lambda_1 \mathcal{L}_{contrast} + \lambda_2 \mathcal{L}_{eqv} + \lambda_3 \mathcal{L}_{unsup} + \lambda_4 \mathcal{L}_{nn}$. We theoretically analyze the effectiveness of our MONA in the very limited label setting (See Section A, available online). We also empirically conduct ablations on different hyperparameters (See Section IV-H).

IV. EXPERIMENTS

In this section, we evaluate our proposed MONA on three popular medical image segmentation datasets under varying labeled ratio settings: the ACDC dataset [92], the LiTS dataset [93], and the MMWHS dataset [94].

A. Datasets

The ACDC dataset was hosted in MICCAI 2017 ACDC challenge [92], 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 use 120, 40 and 40 scans for training, validation, and testing.[3] Note that 1%, 5%, and 10% label ratios denote the ratio of patients. For pre-processing, we adopt the similar setting in [75] 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 [93], which includes 131 contrast-enhanced 3D abdominal CT volumes with expert annotations for two classes (i.e., liver and tumor). Note that 1%, 5%, and 10% label ratios denote the ratio of patients. We use 100 and 31 scans for training, and testing with random order. The splitting details are in the supplementary material. For pre-processing, we adopt the similar setting in [95] 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 [94], 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). Note that 1%, 5%, and 10% label ratios denote the ratio of patients. We use 15 and 5 scans for training and testing with random order. The splitting details are in the supplementary material. 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.

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. Note that 1%, 5%, and 10% label ratios denote the ratio of patients. For ACDC, we adopt the fixed data split [96]. For LiTS and MMWHS, we adopt the random data split with respect to patient.

B. Implementation Details

We implement all the evaluated models using PyTorch library [97]. 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 200 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 [79]. 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 Fig. 6. 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 teacher's input as random rotation, random cropping, horizontal flipping, and strong augmentation to the student's input as random rotation, random cropping, horizontal flipping, random contrast, CutMix [98], brightness changes [99], 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

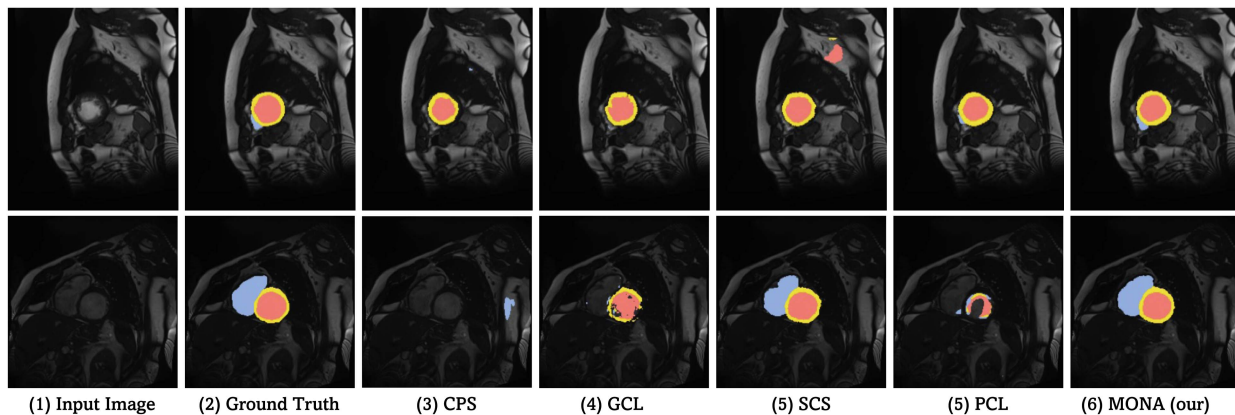| (1) Input Image | (2) Ground Truth | (3) CPS | (4) GCL | (5) SCS | (5) PCL | (6) MONA (our) |

Fig. 5. 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: ▇).
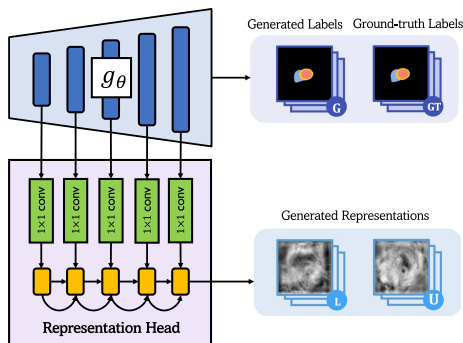

Fig. 6. Overview of the representation head architecture.

the representation head are only used in training, and will be discarded during inference.

C. 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 [92], LiTS [93], MMWHS [94]. We choose 2D UNET [26] as backbone, and compare against SSL methods including EM [88], CCT [89], DAN [68], URPC [90], DCT [62], ICT [91], MT [64], UAMT [51], CPS [49], SIM-CVD [80], MMS [82], SCS [79], GCL [75], and PLC [78]. The upper bound and lower bound method are UNET trained with full/limited supervisions (UNET-F/UNET-L), respectively. We report quantitative comparisons on ACDC and LiTS in Table I.

ACDC: We benchmark performances on ACDC with respect to different labeled ratios (i.e., 1%, 5%, 10%). The following observations can be drawn: *First*, our proposed MONA significantly 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 22.9% and 10.67 in Dice and ASD (i.e., dramatically improving the performance from 59.7% to 82.6%). *Second*, as shown in Fig. 5,

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 I and Fig. 7. The conclusions 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*, our method consistently outperforms all the evaluated SSL methods under different label ratios (i.e., 1%, 5%, 10%). *Third*, as shown in Fig. 7, we observe that MONA is able to produce more accurate results compared to the previous best schemes.

MMWHS. Lastly, we validate MONA on MMWHS, under 1%, 5%, 10% labeled ratios. The results are provided in Table II and Fig. 8. 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 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., GCL) both at the 1% and 5% labeled, yielding 12.3% and 2.8% gains in Dice. We also observe the similar patterns that, MONA performs better or on par with all the other methods at 10% labeled, which again demonstrates the superiority of MONA in extremely limited labeled data regimes.

Overall, we conclude that MONA provides robust performance on all the medical datasets we evaluated, exceeding that of

TABLE I
COMPARISON OF SEGMENTATION PERFORMANCE (DSC[%]/ASD[MM]) ON ACDC AND LiTS UNDER THREE LABELED RATIO SETTINGS (1%, 5%, 10%)

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 [26]	91.5	0.996	91.5	0.996	91.5	0.996	68.5	17.8	68.5	17.8	68.5	17.8
UNET-L	14.5	19.3	51.7	13.1	79.5	2.73	57.0	34.6	60.4	30.4	61.6	28.3
EM [88]	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 [89]	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 [68]	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 [90]	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 [62]	36.0	24.2	58.5	10.8	78.1	2.64	57.6	38.5	60.8	34.4	61.9	31.7
SIMCVD [80]	32.1	20.3	76.1	4.14	79.2	2.21	56.2	32.7	60.5	23.6	61.3	26.0
MMS [82]	32.5	13.6	77.6	3.61	79.4	1.74	56.9	45.6	61.6	55.4	62.5	46.9
ICT [91]	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 [64]	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 [51]	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 [49]	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 [75]	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 [79]	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 [78]	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**	**0.622**	**90.7**	0.864	**64.1**	**20.9**	**67.3**	**16.4**	**69.3**	18.0

The best results are indicated in bold.

(1) Input Image   (2) Ground Truth   (3) CPS   (4) GCL   (5) SCS   (5) PCL   (6) MONA (our)
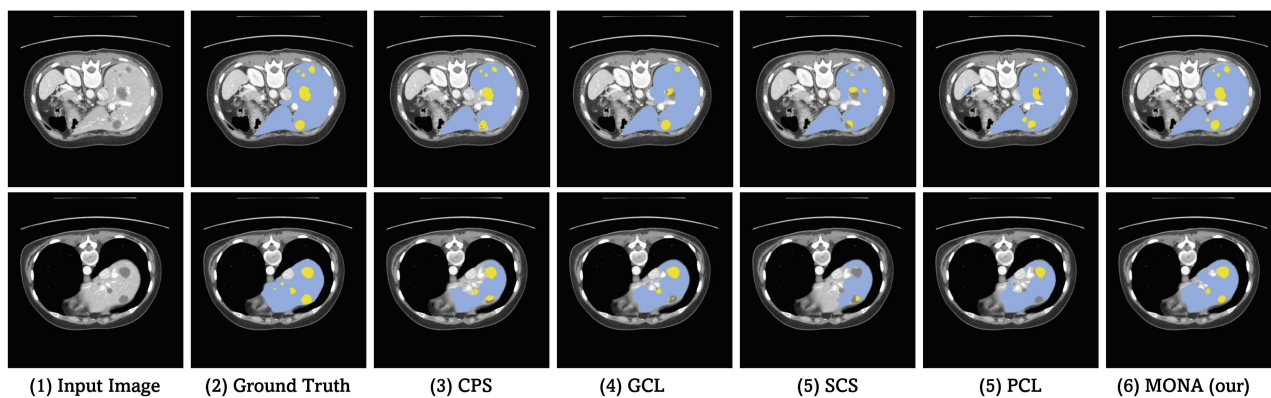
Fig. 7. Visualization of segmentation results on LiTS with 5% labeled ratio. As is shown, MONA consistently produces sharp and accurate object boundaries compared to all other SSL methods. Different anatomical classes are shown in different colors (Liver: ▣; Tumor: ▣).

(1) Input Image   (2) Ground Truth   (3) CPS   (4) GCL   (5) SCS   (5) PCL   (6) MONA (our)
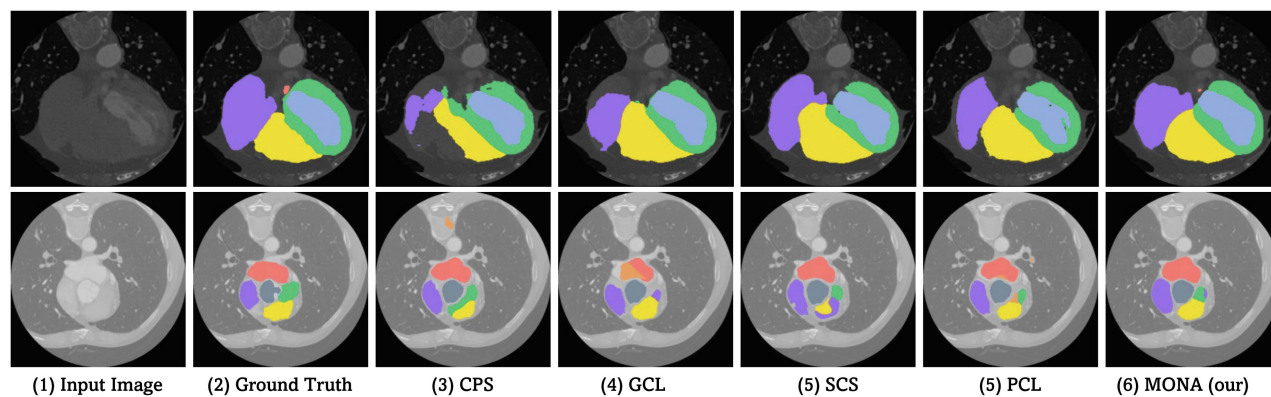
Fig. 8. Visualization of segmentation results on MMWHS with 5% labeled ratio. As is shown, MONA consistently generates more accurate predictions compared to all other SSL methods with a significant performance margin. Different anatomical classes are shown in different colors (LV: ▣; LA: ▣; RV: ▣; RA: ▣; Myo: ▣; PA: ▣).

TABLE II
COMPARISON OF SEGMENTATION PERFORMANCE (DSC[%]/ASD[MM]) ON MMWHS UNDER THREE LABELED RATIO SETTINGS (1%, 5%, 10%)

Method	1% Labeled		5% Labeled		10% Labeled	
	DSC ↑	ASD ↓	DSC ↑	ASD ↓	DSC ↑	ASD ↓
UNET-F [26]	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 [88]	54.5	41.1	80.6	17.3	82.1	15.1
CCT [89]	62.8	27.5	79.0	21.9	79.4	16.3
DAN [68]	52.8	48.4	79.4	22.7	80.2	15.0
URPC [90]	65.7	29.7	73.7	20.5	81.9	12.3
DCT [62]	62.7	27.5	80.8	23.0	82.8	12.4
SIMCVD [80]	64.6	39.5	77.0	20.2	80.3	16.8
MMS [82]	66.2	36.9	80.6	18.4	82.1	16.7
ICT [91]	59.9	32.8	76.5	15.4	82.2	12.0
MT [64]	58.8	35.6	76.5	15.5	79.4	19.8
UAMT [51]	61.1	37.6	76.3	20.9	83.7	14.2
CPS [49]	58.8	33.6	78.3	22.5	82.0	13.1
GCL [75]	71.6	20.3	83.5	**7.41**	86.7	8.76
SCS [79]	71.4	19.3	81.1	11.5	82.6	9.68
PLC [78]	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**

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.

TABLE III
ABLATION ON MODEL COMPONENT: (1) TAILNESS; (2) CONSISTENCY; (3) DIVERSITY, COMPARED TO THE VANILLA AND OUR MONA

Method	Metrics	
	Dice[%] ↑	ASD[mm] ↓
Vanilla	74.2	3.89
w/ tailness	83.1	0.602
w/ consistency	84.2	1.86
w/ diversity	78.2	3.07
w/ tailness + consistency	88.1	0.864
w/ consistency + diversity	80.2	2.11
w/ tailness + diversity	85.0	0.913
● MONA (ours)	**88.8**	**0.622**

the fully-supervised baseline, and outperforming all other SSL methods.

D. Ablation Study

In this subsection, we conduct comprehensive analyses to understand the inner workings of MONA on ACDC under 5% labeled ratio.

E. 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 our CL pre-trained method (i.e., GLCON). 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 III. 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

TABLE IV
ABLATION ON AUGMENTATION STRATEGIES FOR MONA ON THE ACDC AND LITS DATASET UNDER 5% LABELED RATIO

Dataset	Student Aug.	Teacher Aug.	Metrics	
			Dice[%] ↑	ASD[mm] ↓
ACDC	Weak	Weak	86.0	1.02
	Strong	Weak	88.8	0.622
	Weak	Strong	86.4	2.83
	Strong	Strong	88.8	2.07
LiTS	Weak	Weak	62.3	26.5
	Strong	Weak	67.3	16.4
	Weak	Strong	64.3	34.7
	Strong	Strong	66.5	21.1

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 (More results in Section IV-G).

F. 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 [19], [24], [100] 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 teacher's input as random rotation, random cropping, horizontal flipping, and strong augmentation to the student's input as random rotation, random cropping, horizontal flipping, random contrast, CutMix [98], brightness changes [99], morphological changes (diffeomorphic deformations). We summarize the results in Table IV, and list the following observations: (1) **weak augmentations benefits more**: composing the weak augmentation for the teacher model and strong augmentation for the student 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.

TABLE V
ABLATION STUDY OF DIFFERENT CONTRASTIVE LEARNING FRAMEWORKS ON ACDC UNDER THREE LABELED RATIO SETTINGS (1%, 5%, 10%)

Framework	Method	1% Labeled		5% Labeled		10% Labeled	
		DSC ↑	ASD ↓	DSC ↑	ASD ↓	DSC ↑	ASD ↓
only *pretaining*	MoCov2 [7]	38.6	22.4	56.2	17.9	81.0	5.36
	kNN-MoCo [21]	39.5	22.0	58.3	15.7	83.1	7.18
	SimCLR [4]	34.8	24.3	51.7	19.9	80.3	4.16
	BYOL [6]	35.9	7.25	65.9	9.15	85.6	2.51
	ISD [24]	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 [7]	77.7	4.78	85.4	1.52	86.7	1.74
	kNN-MoCo [21]	78.0	4.28	85.9	1.51	86.9	1.61
	SimCLR [4]	75.7	4.33	83.2	2.06	86.1	2.25
	BYOL [6]	77.1	4.84	85.3	2.06	88.1	0.994
	ISD [24]	80.1	3.00	83.8	1.95	88.6	1.20
	● MONA (ours)	**82.6**	**2.03**	**88.8**	**0.622**	**90.7**	**0.864**

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.

G. Generalization Across Contrastive Learning Frameworks

As discussed in Section III-A, 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 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 IV, 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 [7], kNN-MoCo [21], SimCLR [4], BYOL [6], and ISD [24]) with different label settings. More details about these three principles can be found in Section III-B. Of note, MONA can consistently outperform the semi-supervised methods 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 [4], [6], [7], [21], [24] except the epochs number. We train each model in the semi-supervised setting. For labeled data, we follow the same training strategy in Section III-A. 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 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 Section IV-B. For kNN-MoCo [21], given the following ablation study we set the number of neighbors k as 5, and further compare different settings of k in kNN-MoCo [21] 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. Of note, UNET-F is fully supervised.

Comparisons With CL-Based Frameworks: Table V presents the comparisons between our 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 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 [75]. 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%, 33.3%, 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 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 2D medical image segmentation. Particularly, MONA– for the first time – boosts the segmentation performance with 10% labeled ratio over the fully-supervised UNET (UNET-F). From Table I we

TABLE VI
ABLATION STUDY OF DIFFERENT PRINCIPLES ACROSS DIFFERENT CONTRASTIVE LEARNING FRAMEWORKS UNDER VARIOUS LABELED RATIO SETTINGS (1%, 5%, 10%)

Framework	Principle	1% Labeled		5% Labeled		10% Labeled	
		DSC ↑	ASD ↓	DSC ↑	ASD ↓	DSC ↑	ASD ↓
MoCov2 [7]	Vanilla	38.6	22.4	56.2	17.9	81.0	5.36
	tailness	65.0	3.99	81.3	1.13	84.8	1.52
	consistency	70.3	6.88	79.5	3.65	81.9	3.79
	diversity	47.5	10.2	72.2	5.82	83.1	5.46
	tailness + consistency	75.8	5.10	83.8	1.89	85.7	2.81
	consistency + diversity	73.3	6.34	75.4	5.63	82.7	4.39
	tailness + diversity	75.5	5.40	82.4	3.39	85.3	2.49
	tailness + consistency + diversity	77.7	4.78	85.4	1.52	86.7	1.74
kNN-MoCo [21]	Vanilla	39.5	22.0	58.3	15.7	83.1	7.18
	tailness	66.7	3.87	83.7	1.39	86.2	1.17
	consistency	72.2	5.97	81.7	3.13	84.8	3.57
	diversity	50.5	9.53	73.5	5.92	83.5	5.45
	tailness + consistency	76.3	4.51	84.3	2.51	85.7	2.72
	consistency + diversity	72.1	6.45	78.6	5.56	84.6	4.08
	tailness + diversity	75.5	5.75	81.7	3.01	85.6	2.14
	tailness + consistency + diversity	78.0	4.28	85.9	1.51	86.9	1.61
SimCLR [4]	Vanilla	34.8	24.3	51.7	19.9	80.3	4.16
	tailness	61.9	3.52	79.8	1.70	84.5	2.01
	consistency	70.8	5.46	78.1	2.89	84.7	2.24
	diversity	45.9	8.49	68.3	6.46	83.5	3.92
	tailness + consistency	73.0	4.24	83.0	2.43	85.9	2.46
	consistency + diversity	71.1	6.49	75.6	4.47	83.9	3.51
	tailness + diversity	71.9	4.98	81.1	2.92	85.3	2.94
	tailness + consistency + diversity	75.7	4.33	83.2	2.06	86.1	2.25
BYOL [6]	Vanilla	35.9	7.25	65.9	9.15	85.6	2.51
	tailness	64.2	4.26	81.9	1.71	86.4	0.871
	consistency	71.0	5.45	80.2	3.22	87.0	2.08
	diversity	47.5	6.29	70.7	5.48	85.7	2.36
	tailness + consistency	73.7	4.74	83.3	2.01	87.7	1.25
	consistency + diversity	70.9	6.08	76.0	4.55	86.1	1.93
	tailness + diversity	72.2	5.81	82.6	3.12	86.4	1.33
	tailness + consistency + diversity	77.1	4.84	85.3	2.06	88.1	0.994
ISD [24]	Vanilla	45.8	17.2	71.0	4.29	85.3	2.97
	tailness	71.8	2.80	79.2	1.47	87.1	1.02
	consistency	78.8	3.98	80.2	2.90	87.3	1.94
	diversity	54.5	8.03	77.1	6.90	86.2	2.58
	tailness + consistency	79.6	2.99	83.0	1.93	88.2	1.24
	consistency + diversity	75.1	4.72	77.8	3.65	86.5	2.45
	tailness + diversity	74.8	7.98	82.3	2.02	87.2	1.35
	tailness + consistency + diversity	80.1	3.00	83.8	1.95	88.6	1.20
MONA (ours)	Vanilla	49.3	7.11	74.2	3.89	86.5	1.92
	tailness	75.1	1.83	83.1	0.602	87.8	0.577
	consistency	81.5	2.78	84.2	1.86	88.4	1.33
	diversity	62.8	3.97	78.2	3.07	86.6	1.88
	tailness + consistency	81.2	2.19	88.1	0.864	90.1	0.966
	consistency + diversity	81.8	3.29	80.2	2.11	86.9	1.67
	tailness + diversity	78.6	3.33	85.0	0.913	89.5	0.673
	tailness + consistency + diversity	82.6	2.03	88.8	0.622	90.7	0.864

Experiments are conducted on ACDC using UNET [26] as the backbone with three independent runs. Here we report the segmentation performance in terms of DSC[%] and ASD[mm]. On all three labeled settings, incorporating our methods (i.e., tailness, consistency, and diversity) consistently achieve superior model robustness gains across different state-of-the-art CL frameworks.

see that on LiTS with 10% labeled ratio, MONA outperforms UNET-F by 0.8 in terms of DSC (69.3 vs 68.5). From Table II, MONA outperforms UNET-F on MMWHS by 1.8 in terms of DSC (87.6 vs 85.8). Tables I and II also show that MONA significantly outperforms all the other semi-supervised methods by a large margin. In summary, our 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 [4], [6], [7], [21], [24] adopting the two-branch design, demonstrating that these concepts consistently help the model yield extra performance boosts for them all.

Generalization Across CL Frameworks: As demonstrated in Table VI, incorporating *tailness*, *consistency*, and *diversity* have obviously superior performance boosts, which is aligned with consistent observations with Section IV-D can be drawn.
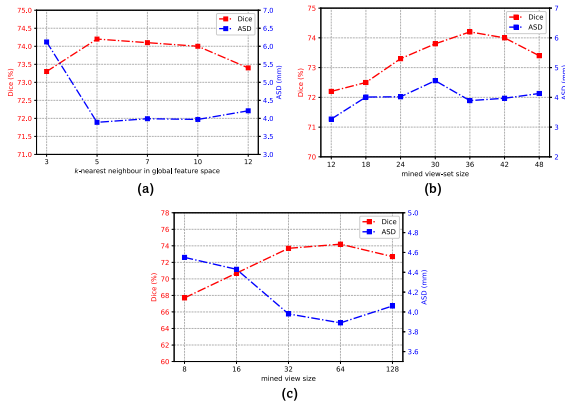
Fig. 9. Effects of k-nearest neighbour in global feature space, mined view-set size, and mined view patch size. We report Dice and ASD of GLCON on the ACDC dataset at the 5% labeled ratio. All the experiments are run with three different random seeds.
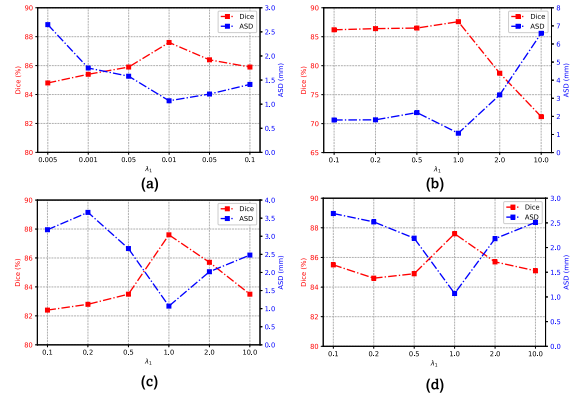

Fig. 10. Effects of hyperparameters $\lambda_1, \lambda_2, \lambda_3, \lambda_4$. We report Dice and ASD of MONA on the ACDC dataset at the 5% labeled ratio. All the experiments are run with three different random seeds.

This suggests that these three principles can serve as desirable properties for medical image segmentation in both supervised and unsupervised settings.

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 [21]. 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 [21] (More analysis on data augmentation can be found in Section IV-D), 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 $k \in \{3, 5, 7, 10, 12\}$. Fig. 9(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 Fig. 9(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 $\in \{8, 16, 32, 64, 128\}$. Fig. 9(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.

H. 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 $\mathcal{L}_{total}$: Proper choices of hyperparameters in total loss $\mathcal{L}_{\text{total}}$ (See Section III-B) 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 Fig. 10, 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 $\delta_\theta \in$
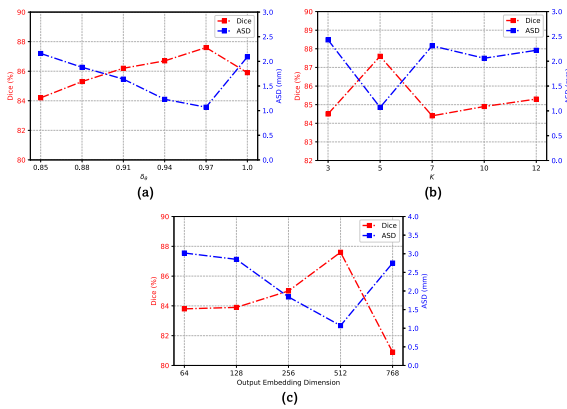
Fig. 11. Effects of confidence threshold δ_θ, K-nearest neighbour constraint, and output embedding dimension. We report Dice and ASD of MONA on the ACDC dataset at the 5% labeled ratio. All the experiments are run with three different random seeds.

$\{0.85, 0.88, 0.91, 0.94, 0.97, 1.0\}$. Fig. 11(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 $K \in \{3, 5, 7, 10, 12\}$. Fig. 11(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.

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 $\in \{64, 128, 256, 512, 768\}$. Fig. 11(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.

V. CONCLUSION

In this paper, we have presented MONA, a semi-supervised contrastive learning method for 2D 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. Both empirical and theorical studies show 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 superiority 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.

REFERENCES

[1] Z. Wu, Y. Xiong, X. Y. Stella, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3733–3742.

[2] A. V. D. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv:1807.03748*.

[3] R. D. Hjelm et al., "Learning deep representations by mutual information estimation and maximization," in *Proc. Int. Conf. Learn. Representations*, 2019.

[4] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[5] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9726–9735 .

[6] J.-B. Grill et al., "Bootstrap your own latent-a new approach to self-supervised learning," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 21271–21284.

[7] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," 2020, *arXiv: 2003.04297*.

[8] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 9912–9924.

[9] Q. Quan et al., "Information-guided pixel augmentation for pixel-wise contrastive learning," 2022, *arXiv:2211.07118*.

[10] J. Li, Q. Quan, and S. K. Zhou, "MixCL: Pixel label matters to contrastive learning," 2022, *arXiv:2203.02114*.

[11] S. Arora, H. Khandeparkar, M. Khodak, O. Plevrakis, and N. Saunshi, "A theoretical analysis of contrastive unsupervised representation learning," 2019, *arXiv: 1902.09229*.

[12] C.-Y. Chuang, J. Robinson, Y.-C. Lin, A. Torralba, and S. Jegelka, "Debiased contrastive learning," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 8765–8775.

[13] J. Li, P. Zhou, C. Xiong, and S. Hoi, "Prototypical contrastive learning of unsupervised representations," in *Proc. Int. Conf. Learn. Representations*, 2021.

[14] B. Kang, Y. Li, S. Xie, Z. Yuan, and J. Feng, "Exploring balanced feature spaces for representation learning," in *Proc. Int. Conf. Learn. Representations*, 2020.

[15] X. Zhu, D. Anguelov, and D. Ramanan, "Capturing long-tail distributions of object subcategories," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 915–922.

[16] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 9268–9277.

[17] Y. Yang and Z. Xu, "Rethinking the value of labels for improving class-imbalanced learning," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 19290–19301.

[18] Z. Jiang, T. Chen, T. Chen, and Z. Wang, "Improving contrastive learning on imbalanced data via open-world sampling," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 5997–6009.

[19] M. Zheng et al., "ReSSL: Relational self-supervised learning with weak augmentation," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 2543–2555.

[20] M. Azabou et al., "Mine your own view: Self-supervised learning through across-sample prediction," 2021, *arXiv:2102.10106*.

[21] W. Van Gansbeke, S. Vandenhende, S. Georgoulis, and L. V. Gool, "Revisiting contrastive methods for unsupervised learning of visual representations," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 16238–16250.

[22] S. Liu, S. Zhi, E. Johns, and A. J. Davison, "Bootstrapping semantic segmentation with regional contrast," 2021, *arXiv:2104.04465*.

[23] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2006, pp. 1735–1742.

[24] A. Tejankar, S. A. Koohpayegani, V. Pillai, P. Favaro, and H. Pirsiavash, "ISD: Self-supervised learning by iterative similarity distillation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 9589–9598.

[25] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.

[26] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, 2015, pp. 234–241.

[27] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[29] A. Vaswani et al., "Attention is all you need," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[30] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2020.

[31] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. Int. Conf. 3D Vis.*, 2016, pp. 565–571.

[32] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[33] M. Z. Alom, M. Hasan, C. Yakopcic, T. M. Taha, and V. K. Asari, "Recurrent residual convolutional neural network based on U-Net (R2U-Net) for medical image segmentation," 2018, *arXiv: 1802.06955*.

[34] O. Oktay et al., "Attention U-net: Learning where to look for the pancreas," 2018, *arXiv: 1804.03999*.

[35] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 833–851.

[36] J. Chen et al., "TransUNet: Transformers make strong encoders for medical image segmentation," 2021, *arXiv:2102.04306*.

[37] H. Cao et al., "Swin-unet: Unet-like pure transformer for medical image segmentation," 2021, *arXiv:2105.05537*.

[38] Y. Xie, J. Zhang, C. Shen, and Y. Xia, "CoTr: Efficiently bridging CNN and transformer for 3d medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, 2021, pp. 171–180.

[39] A. Hatamizadeh et al., "UNETR: Transformers for 3D medical image segmentation," 2021, *arXiv:2103.10504*.

[40] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, and V. M. Patel, "Medical transformer: Gated axial-attention for medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, 2021, pp. 36–46.

[41] C. You et al., "Class-aware generative adversarial transformers for medical image segmentation," 2022, *arXiv:2201.10737*.

[42] J. Dai et al., "Deformable convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 764–773.

[43] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2999–3007.

[44] Y. Xue et al., "Shape-aware organ segmentation by predicting signed distance maps," 2019, *arXiv: 1912.03849*.

[45] G. Shi, L. Xiao, Y. Chen, and S. K. Zhou, "Marginal loss and exclusion loss for partially supervised multi-organ segmentation," *Med. Image Anal.*, vol. 70, 2021, Art. no. 101979.

[46] D.-H. Lee et al., "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Proc. Workshop Challenges Representation Learn.*, 2013, pp. 896–902.

[47] W. Bai et al., "Semi-supervised learning for network-based cardiac MR image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, 2017, pp. 253–260.

[48] D.-P. Fan et al., "Inf-net: Automatic COVID-19 lung infection segmentation from CT images," *IEEE Trans. Med. Imag.*, vol. 39, no. 8, pp. 2626–2637, Aug. 2020.

[49] X. Chen, Y. Yuan, G. Zeng, and J. Wang, "Semi-supervised semantic segmentation with cross pseudo supervision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 2613–2622.

[50] I. Nassar, M. Hayat, E. Abbasnejad, H. Rezatofighi, and G. Haffari, "ProtoCon: Pseudo-label refinement via online clustering and prototypical consistency for efficient semi-supervised learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 11641–11650.

[51] L. Yu, S. Wang, X. Li, C.-W. Fu, and P.-A. Heng, "Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, 2019, pp. 605–613.

[52] S. Graham et al., "Mild-net: Minimal information loss dilated network for gland instance segmentation in colon histology images," *Med. Image Anal.*, vol. 52, pp. 199–211, 2019.

[53] X. Cao, H. Chen, Y. Li, Y. Peng, S. Wang, and L. Cheng, "Uncertainty aware temporal-ensembling model for semi-supervised abus mass segmentation," *IEEE Trans. Med. Imag.*, vol. 40, no. 1, pp. 431–443, Jan. 2021.

[54] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, "Weight uncertainty in neural network," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1613–1622.

[55] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1050–1059.

[56] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 5574–5584.

[57] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, "Self-training with noisy student improves imagenet classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10684–10695 .

[58] G. Bortsova, F. Dubost, L. Hogeweg, I. Katramados, and M. de Bruijne, "Semi-supervised medical image segmentation via learning consistency under transformations," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, 2019, pp. 810–818.

[59] W. Cui et al., "Semi-supervised brain lesion segmentation with an adapted mean teacher model," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, 2019, pp. 554–565.

[60] G. Fotedar, N. Tajbakhsh, S. Ananth, and X. Ding, "Extreme consistency: Overcoming annotation scarcity and domain shifts," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, 2020, pp. 699–709.

[61] M. Sajjadi, M. Javanmardi, and T. Tasdizen, "Regularization with stochastic transformations and perturbations for deep semi-supervised learning," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 1163–1171.

[62] S. Qiao, W. Shen, Z. Zhang, B. Wang, and A. Yuille, "Deep co-training for semi-supervised image recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 142–159.

[63] Y. Zhou et al., "Semi-supervised 3D abdominal multi-organ segmentation via deep multi-planar co-training," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2019, pp. 121–140.

[64] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 1195–1204.

[65] Z. Zhao, F. Zhou, Z. Zeng, C. Guan, and S. K. Zhou, "Meta-hallucinator: Towards few-shot cross-modality cardiac image segmentation," in *Proc. 25th Int. Conf. Med. Image Comput. Comput. Assist. Intervention*, Singapore, 2022, pp. 128–139.

[66] X. Li, L. Yu, H. Chen, C.-W. Fu, L. Xing, and P.-A. Heng, "Transformation-consistent self-ensembling model for semisupervised medical image segmentation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 2, pp. 523–534, Feb. 2021.

[67] S. Reiß, C. Seibold, A. Freytag, E. Rodner, and R. Stiefelhagen, "Every annotation counts: Multi-label deep supervision for medical image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 9532–9542.

[68] Y. Zhang, L. Yang, J. Chen, M. Fredericksen, D. P. Hughes, and D. Z. Chen, "Deep adversarial networks for biomedical image segmentation utilizing unannotated images," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, 2017, pp. 408–416.

[69] D. Nie, Y. Gao, L. Wang, and D. Shen, "ASDNet: Attention based semi-supervised deep networks for medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, 2018, pp. 370–378.

[70] Z. Zhang, L. Yang, and Y. Zheng, "Translating and segmenting multimodal medical volumes with cycle-and shape-consistency generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 9242–9251.

[71] H. Zheng etal, "Semi-supervised segmentation of liver using adversarial learning with deep atlas prior," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, 2019, pp. 148–156.

[72] S. Li, C. Zhang, and X. He, "Shape-aware semi-supervised 3D semantic segmentation for medical images," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, 2020, pp. 552–561.

[73] G. Valvano, A. Leo, and S. A. Tsaftaris, "Learning to segment from scribbles using multi-scale adversarial attention gates," *IEEE Trans. Med. Imag.*, vol. 40, no. 8, pp. 1990–2001, Aug. 2021.

[74] Y. Grandvalet and Y. Bengio, "Semi-supervised learning by entropy minimization," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2004, pp. 281–296 .

[75] K. Chaitanya, E. Erdil, N. Karani, and E. Konukoglu, "Contrastive learning of global and local features for medical image segmentation with limited annotations," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 12546–12558.

[76] D. Dwibedi, Y. Aytar, J. Tompson, P. Sermanet, and A. Zisserman, "With a little help from my friends: Nearest-neighbor contrastive learning of visual representations," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 9568–9577.

[77] C. You, R. Zhao, L. Staib, and J. S. Duncan, "Momentum contrastive voxel-wise representation learning for semi-supervised volumetric medical image segmentation," 2021, *arXiv:2105.07059*.

[78] K. Chaitanya, E. Erdil, N. Karani, and E. Konukoglu, "Local contrastive loss with pseudo-label based self-training for semi-supervised medical image segmentation," 2021, *arXiv:2112.09645*.

[79] X. Hu, D. Zeng, X. Xu, and Y. Shi, "Semi-supervised contrastive learning for label-efficient medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, 2021, pp. 481–490.

[80] C. You, Y. Zhou, R. Zhao, L. Staib, and J. S. Duncan, "SimCVD: Simple contrastive voxel-wise representation distillation for semi-supervised medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 41, no. 9, pp. 2228–2237, Sep. 2022.

[81] S. Zhang, J. Zhang, B. Tian, T. Lukasiewicz, and Z. Xu, "Multimodal contrastive mutual learning and pseudo-label re-learning for semi-supervised medical image segmentation," *Med. Image Anal.*, vol. 83, 2023, Art. no. 102656.

[82] A. Lou, K. Tawfik, X. Yao, Z. Liu, and J. Noble, "Min-max similarity: A contrastive semi-supervised deep learning network for surgical tools segmentation," *IEEE Trans. Med. Imag.*, vol. 42, no. 10, pp. 2832–2841, Oct. 2023.

[83] R. Dangovski et al., "Equivariant contrastive learning," in *Proc. Int. Conf. Learn. Representations*, 2022.

[84] A. Galdran, G. Carneiro, and M. A. González Ballester, "Balanced-mixup for highly imbalanced medical image classification," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, 2021, pp. 323–333.

[85] K. Yan et al., "SAM: Self-supervised learning of pixel-wise anatomical embeddings in radiological images," *IEEE Trans. Med. Imag.*, vol. 41, no. 10, pp. 2658–2669, Oct. 2022.

[86] A. G. Roy et al., "Does your dermatology classifier know what it doesn't know? Detecting the long-tail of unseen conditions," *Med. Image Anal.*, vol. 75, 2022, Art. no. 102274.

[87] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 936–944.

[88] T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. Pérez, "Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2517–2526.

[89] Y. Ouali, C. Hudelot, and M. Tami, "Semi-supervised semantic segmentation with cross-consistency training," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 12671–12681.

[90] X. Luo et al., "Efficient semi-supervised gross target volume of nasopharyngeal carcinoma segmentation via uncertainty rectified pyramid consistency," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, 2021, pp. 318–329.

[91] V. Verma, K. Kawaguchi, A. Lamb, J. Kannala, Y. Bengio, and D. Lopez-Paz, "Interpolation consistency training for semi-supervised learning," in *Proc. Int. Joint Conf. Artif. Intell.*, 2019, pp. 3635–3641.

[92] O. Bernard et al., "Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved?," *IEEE Trans. Med. Imag.*, vol. 37, no. 11, pp. 2514–2525, Nov. 2018.

[93] P. Bilic et al., "The liver tumor segmentation benchmark (LiTS)," 2019, *arXiv: 1901.04056*.

[94] X. Zhuang and J. Shen, "Multi-scale patch and multi-modality atlases for whole heart segmentation of MRI," *Med. Image Anal.*, vol. 31, pp. 77–87, 2016.

[95] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, "H-denseunet: Hybrid densely connected Unet for liver and tumor segmentation from CT volumes," *IEEE Trans. Med. Imag.*, vol. 37, no. 12, pp. 2663–2674, Dec. 2018.

[96] Y. Wu, Z. Wu, Q. Wu, Z. Ge, and J. Cai, "Exploring smoothness and class-separation for semi-supervised medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, 2022, pp. 34–43.

[97] A. Paszke et al., "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2019, pp. 8024–8035.

[98] G. French, S. Laine, T. Aila, M. Mackiewicz, and G. Finlayson, "Semi-supervised semantic segmentation needs strong, varied perturbations," in *Proc. Brit. Mach. Vis. Conf.*, 2020.

[99] F. Perez, C. Vasconcelos, S. Avila, and E. 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*. Berlin, Germany: Springer, 2018, pp. 303–311.

[100] K. Sohn et al., "FixMatc: Simplifying semi-supervised learning with consistency and confidence," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 596–608.

[101] P. L. Bartlett and S. Mendelson, "Rademacher and gaussian complexities: Risk bounds and structural results," *J. Mach. Learn. Res.*, vol. 3, no. Nov, pp. 463–482, 2002.

[102] T. Furlanello, Z. Lipton, M. Tschannen, L. Itti, and A. Anandkumar, "Born again neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1602–1611.

[103] Z. Zhang and M. Sabuncu, "Self-distillation as instance-specific label smoothing," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 2184–2195.

[104] H. Mobahi, M. Farajtabar, and P. Bartlett, "Self-distillation amplifies regularization in hilbert space," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 3351–3361.

Chenyu You received the BS degree from Rensselaer Polytechnic Institute and the MSc degree from Stanford University. He is currently working toward the PhD degree with Yale University. His research interests include broadly in the area of machine learning theory and algorithms intersecting the fields of computer & medical vision, natural language processing, and signal processing. Previously, he has published several works in NeurIPS, ICLR, CVPR, MICCAI, ACL, EMNLP, NAACL, AAAI, IJCAI, and KDD, etc. He has been awarded as the Outstanding/Distinguished Reviewer of CVPR, MICCAI, *IEEE Transactions on Medical Imaging* (TMI), and Medical Physics. He serves as an area chair at MICCAI, and has co-led workshops such as foundation model on healthcare (ICML).

Weicheng Dai received the BS degree from Southeast University and the MSc degree from New York University. He is a postgraduate associate with Yale University. His research interests include computer vision, medical image analysis, natural language processing, and theoretical machine learning. He has published papers at top-tier conferences in MICCAI and IPMI, etc. He has served as reviewers for *IEEE Transactions on Medical Imaging*, MICCAI, AAAI, and IJCAI.

Fenglin Liu is currently working toward the PhD degree with the University of Oxford. His research interests include natural language processing (NLP), especially vision-and-language, machine learning, and their applications to clinical, i.e., clinical NLP. He has published papers at top-tier journals and conferences, e.g., *IEEE Transactions on Pattern Analysis and Machine Intelligence*, NeurIPS, CVPR, ACL, EMNLP, NAACL. He has served as a senior program committee member for IJCAI and was awarded as the Distinguished/Outstanding Reviewer of CVPR, AAAI, and IJCAI.

Yifei Min received the BS degree from the University of Hong Kong and MA degree from the University of Pennsylvania, respectively. He is currently working toward the PhD degree in statistics and data science with Yale University. He has a broad research interest include the area of machine learning theory with a focus on online learning, reinforcement learning, and medical imaging analysis. Previously, several of his work have been published in NeurIPS, ICLR, ICML, UAI, AISTATS, etc.

Nicha C. Dvornek received the BS degree in biomedical engineering from Johns Hopkins University and the MS, MPhil, and PhD degrees in engineering and applied science from Yale University. She is an assistant professor with Radiology & Biomedical Imaging and Biomedical Engineering, Yale University. She completed postdoctoral training in diagnostic radiology wth Yale and as a T32 fellow with the Yale Child Study Center. Her recent work focuses on the development and application of novel machine learning approaches to medical image processing and analysis, with applications spanning from neurodevelopmental disorders to cancer. She and her team's work has been recognized by multiple paper and poster awards at international meetings. She is a member of the MICCAI society and on the board of the Women in MICCAI. She serves as an associate editor for Computerized Medical Imaging and Graphics and the *Journal of Medical Imaging*, an area chair for MICCAI and MIDL conferences, and a reviewer for multiple journals and conferences such as Medical Image Analysis, *IEEE Transactions on Medical Imaging*, NeurIPS, and CVPR.

Xiaoxiao Li received the PhD degree from Yale University, New Haven, CT, USA, in 2020. She was a postdoctoral research fellow with the Computer Science Department, Princeton University, Princeton, NJ, USA. Since August 2021, she has been an assistant professor with the Department of Electrical and Computer Engineering (ECE), University of British Columbia (UBC), Vancouver, BC, Canada. In the last few years, she has more than 30 papers published in leading machine learning conferences and journals, including NeurIPS, ICML, ICLR, MICCAL, IPML, BMVC, *IEEE Transactions On Medical Lmaging*, and *Medical Image Analysis*. Her research interests include range across the interdisciplinary fields of deep learning and biomedical data analysis, aiming to improve the trustworthiness of artificial intelligence (AI) systems for health care. She work has been recognized with several best paper awards at international conferences.

David A. Clifton is the Royal Academy of Engineering Chair of Clinical Machine Learning with the University of Oxford, and OCC fellow in AI & Machine Learning with Reuben College, Oxford. He was the first AI scientist to be appointed to an NIHR Research Professorship, which is the U.K. medical research community's "flagship Chair programme". He is a fellow of the Alan Turing Institute, research fellow of the Royal Academy of Engineering, Visiting Chair in AI for Healthcare with the University of Manchester, and a fellow of Fudan University, China. He studied Information Engineering at Oxford's Department of Engineering Science, supervised by Prof. Lionel Tarassenko CBE, Chair of Electrical Engineering. His research focuses on the development of machine learning for tracking the health of complex systems. His previous research resulted in patented systems for jet-engine health monitoring, used with the engines of the Airbus A380, the Boeing 787 "Dreamliner", and the Eurofighter Typhoon. Since graduating from his DPhil in 2009, he has focused mostly on the development of AI-based methods for healthcare. Patents arising from this collaborative research have been commercialised via university spin-out companies OBS Medical, Oxehealth, and Sensyne Health, in addition to collaboration with multinational industrial bodies. He was awarded a Grand Challenge award from the U.K. Engineering and Physical Sciences Research Council, which is an EPSRC Fellowship that provides long-term strategic support for "future leaders in healthcare". His research has been awarded over 35 academic prizes; in 2018, he was joint winner of the inaugural "Vice-Chancellor's Innovation Prize", which identifies the best interdisciplinary research across the entirety of the University of Oxford.

Lawrence Staib received the BA degree in physics from Cornell University, and the PhD degree in engineering and applied science from Yale University. He is professor of radiology & biomedical imaging, biomedical engineering, and electrical engineering with Yale University. He is director of undergraduate studies in biomedical engineering with Yale. He is a fellow of the American Institute for Medical and Biological Engineering and of the Medical Image Computing and Computer-Assisted Intervention Society. He is a distinguished investigator of the Academy for Radiology & Biomedical Imaging Research. He is a member of the editorial board of Medical Image Analysis and associate editor of *IEEE Transactions on Biomedical Engineering*. His research interests include medical image analysis and machine learning.

James S. Duncan (Life Fellow, IEEE) received the BS degree in electrical engineering with honors from Lafayette College, in 1973, the MS degree in electrical engineering from the University of California, Los Angeles, in 1975, and the PhD degree in electrical engineering from the University of Southern California, in 1982. He is the Ebenezer K. Hunt professor of biomedical engineering and a professor of radiology & biomedical engineering, electrical engineering and statistics & data science with Yale University. He is currently the chair of the Department of Biomedical Engineering. His research efforts have been in the areas of computer vision, image processing, and medical imaging, with an emphasis on biomedical image analysis. These efforts have included the segmentation of deformable structure from 3D image data, the tracking of non-rigid motion/deformation from spatiotemporal images, and the development strategies for image-guided intervention/surgery. Most recently, his efforts have focused on the development of image analysis strategies based on the integration of data-driven machine learning and physical model-based approaches. He has published over 325 peer-reviewed articles in these areas and has been the Principal investigator on a number of peer-reviewed grants from both the National Institutes of Health and the National Science Foundation over the past 40 years. He is a fellow of the Institute of Electrical and Electronic Engineers (IEEE), of the American Institute for Medical and Biological Engineering (AIMBE) and of the Medical Image Computing and Computer Assisted Intervention (MICCAI) Society. In 2012, he was elected to the Council of Distinguished Investigators, Academy of Radiology Research. In 2014 he was elected to the Connecticut Academy of Science & Engineering. He currently serves as the co-editor-in-chief of Medical Image Analysis, and has served on the editorial boards of the *IEEE Transactions on Medical Imaging*, the *Journal of Mathematical Imaging and Vision* and the *Proceedings of the IEEE*. He is a past president of the MICCAI Society and in 2017 received the MICCAI Society's Enduring Impact Award.