Evaluating semi-supervision methods for medical image segmentation: applications in cardiac magnetic resonance imaging

Sarah M. Hooper[a,*] **Sen Wu**[b] **Rhodri H. Davies**[c,d,e] **Anish Bhuva**[c,d] **Erik B. Schelbert**[f,g,h] **James C. Moon**[c,d] **Peter Kellman**[i] **Hui Xue**[i] **Curtis Langlotz**[j] **and Christopher Ré**[b]

[a]Stanford University, Department of Electrical Engineering, Stanford, California, United States
[b]Stanford University, Department of Computer Science, Stanford, California, United States
[c]Barts Health NHS Trust, Barts Heart Centre, London, United Kingdom
[d]University of College London, Institute of Cardiovascular Sciences, London, United Kingdom
[e]University of College London, MRC Centre for Lifelong Health and Ageing, London, United Kingdom
[f]United Hospital, St. Paul, Minnesota, and Abbott Northwestern Hospital, Minneapolis Heart Institute, Minneapolis, Minnesota, United States
[g]UPMC Cardiovascular Magnetic Resonance Center, UPMC, Pittsburgh, Pennsylvania, United States
[h]University of Pittsburgh School of Medicine, Department of Medicine, Pittsburgh, Pennsylvania, United States
[i]National Heart, Lung and Blood Institute, National Institutes of Health, Bethesda, Maryland, United States
[j]Stanford University, Department of Radiology, Department of Biomedical Informatics, Stanford, California, United States

Abstract

Purpose: Neural networks have potential to automate medical image segmentation but require expensive labeling efforts. While methods have been proposed to reduce the labeling burden, most have not been thoroughly evaluated on large, clinical datasets or clinical tasks. We propose a method to train segmentation networks with limited labeled data and focus on thorough network evaluation.

Approach: We propose a semi-supervised method that leverages data augmentation, consistency regularization, and pseudolabeling and train four cardiac magnetic resonance (MR) segmentation networks. We evaluate the models on multiinstitutional, multiscanner, multidisease cardiac MR datasets using five cardiac functional biomarkers, which are compared to an expert's measurements using Lin's concordance correlation coefficient (CCC), the within-subject coefficient of variation (CV), and the Dice coefficient.

Results: The semi-supervised networks achieve strong agreement using Lin's CCC (>0.8), CV similar to an expert, and strong generalization performance. We compare the error modes of the semi-supervised networks against fully supervised networks. We evaluate semi-supervised model performance as a function of labeled training data and with different types of model supervision, showing that a model trained with 100 labeled image slices can achieve a Dice coefficient within 1.10% of a network trained with 16,000+ labeled image slices.

Conclusion: We evaluate semi-supervision for medical image segmentation using heterogeneous datasets and clinical metrics. As methods for training models with little labeled data become more common, knowledge about how they perform on clinical tasks, how they fail, and how they perform with different amounts of labeled data is useful to model developers and users.

 [DOI: 10.1117/1.JMI.10.2.024007]

*Address all correspondence to Sarah M. Hooper, smhooper@stanford.edu

Keywords: deep learning; semi-supervision; cardiac magnetic resonance; segmentation.

Paper 22198GR received Aug. 2, 2022; accepted for publication Feb. 27, 2023; published online Mar. 30, 2023.

1 Introduction

Deep neural networks (DNNs) offer state-of-the-art performance for automated medical image segmentation,[1–3] decreasing the time required to analyze images and the intra- and inter-observer variation.[2,4] However, labeling data to train these models is labor intensive and requires the expertise of well-trained medical specialists, impeding the creation of large, diverse labeled training datasets and leading to delays in developing clinical machine learning models. In contrast, most healthcare facilities have abundant unlabeled medical imaging data readily available. In response, there is high interest in reducing the image labeling burden by leveraging unlabeled data to train models. Recently, many promising semi-[5–8] and self-supervision[9,10] methods have been proposed to accomplish this goal.

However, most prior work in self- and semi-supervised medical image segmentation has focused on evaluating image processing metrics like the Dice coefficient instead of clinically used metrics, or has used small (e.g., 40 to 100 patients including train and test splits) or single-institution datasets to train and evaluate their networks. This narrow evaluation leaves gaps in our understanding of how segmentation models trained with limited labeled data perform in important clinical settings. For example, without evaluating semi-supervised networks on clinical biomarkers, it is unknown if semi-supervised segmentation methods lead to accurate clinical measurements. By working with small datasets, it is unknown how networks trained with limited labeled data compare to networks trained with abundant labeled data, particularly on difficult edge cases. Similarly, by training and evaluating on single institution datasets, it is unclear if segmentation networks trained with limited labeled data generalize to large, multi-institutional datasets with different patient cohorts and equipment as seen during deployment.

In this work, we aim to address these missing evaluations, focusing on how medical image segmentation networks trained with very limited labeled data perform on important clinical tasks. We focus on training and evaluating segmentation networks for cardiac magnetic resonance (CMR) imaging, an area in which segmentation plays a key and common role.[11,12] To perform these evaluations, we use multiinstitutional, multiscanner, multidisease datasets, which are more reflective of the data heterogeneity seen during network deployment.

We start by defining our semi-supervised training pipeline. This proposed pipeline leverages multiple approaches to model supervision—including data augmentation (DA), consistency regularization, and pseudolabeling (PL)—into one training procedure, enabling us to leverage and evaluate multiple techniques for training medical image segmentation models with limited labels. We use this pipeline to train semi-supervised segmentation models with 100 labeled image slices on four CMR segmentation tasks.

We next evaluate these semi-supervised networks on clinically important metrics by using the predicted segmentation masks to derive five cardiac functional parameters, which quantitate the heart's function or structure and are relied upon to inform patient care. We evaluate the accuracy of these functional parameters by comparing the values computed by the semi-supervised networks against the values computed by a human expert using Lin's concordance correlation coefficient (CCC). Next, we compute the within-subject coefficient of variation (CV) of the semi-supervised networks and compare against the CV of a human expert. Finally, we assess if the semi-supervised networks generalize to data collected from an independent hospital. For all of these evaluations, we further compare the semi-supervised networks against naïve models (i.e., networks trained with only 100 labeled image slices and no semi-supervision methods) to assess if semi-supervision leads to performance improvements on clinically meaningful metrics.

Finally, we dig into the proposed training procedure to understand how each component of the training pipeline contributes to network performance. First, we assess how networks trained

with limited data compare to networks trained with abundant labeled data by sweeping the number of labeled training images from 10 to 16,812 images. We further examine how each different source of supervision in the training pipeline—namely, DA, consistency regularization, and PL—impacts model performance and we assess if these sources of supervision can be used together synergistically.

In summary, our contributions include:

- A semi-supervised pipeline that incorporates multiple approaches to training medical image segmentation networks with little labeled data. By folding multiple approaches to limited-label training into the same framework, we achieve high performance and provide a flexible, end-to-end system to train segmentation networks with limited labeled data. This pipeline will be available on GitHub with tutorials for how to train segmentation networks on custom datasets.
- Evaluation of the proposed semi-supervised segmentation pipeline on diverse, multicenter clinical datasets for CMR segmentation. We evaluate five clinically relevant imaging biomarkers that assess cardiac function and evaluate the network's ability to generalize to external datasets.
- Ablation studies to assess how the error modes of models trained with limited data differ from those trained with abundant data, which different methods of training with limited labeled data most impact performance, and how each method performs with different amounts of labeled data.

2 Related Work

2.1 *CMR Segmentation*

Much of the prior work in deep learning-based CMR segmentation has focused on short-axis (SAX) cine images.[13] For example, a network architecture has been proposed to incorporate 2D and volumetric information for segmenting the left ventricle (LV) endocardium, myocardium, and right ventricle using a dataset of 150 SAX cine stacks.[14] A variational autoencoder can enforce anatomically plausible outputs from a segmentation network.[15] Additional fully supervised SAX cine segmentation networks are discussed in Refs. 13 and 16–18. Comparatively less work has been done segmenting structures in long-axis (LAX) cine using deep learning. A dataset of 63 patients was used to segment the LV endocardium and myocardium in 2-chamber LAX (2CH LAX), 4-chamber LAX (4CH LAX), and SAX cines as well as the atria on LAX cines, where all tasks were performed using the same network.[19] 90 patients were used in a different study training neural networks to segment the LV and RV on LAX cines.[20] The large UK Biobank dataset of 4875 subjects was used to train networks to segment the atria on LAX CMR.[21] Other examples of deep learning-based CMR segmentation include work on CMR perfusion images,[22] T1 maps,[23] and late gadolinium enhancement.[24] The above approaches rely on fully supervised learning. We study CMR in our work because we know that fully supervised networks perform well in this setting and large labeled datasets are available for comparison; we sought to understand whether networks trained with little labeled data could attain similar performance, and if not, what the error modes were. We build on these past approaches by training models with much less labeled data, and the training approach we use in this work can be used for new applications that do not have public networks/datasets.

2.2 *Reducing the Amount of Required Labeled Data*

Collecting labeled data is a common bottleneck in machine learning pipelines. A great deal of prior work has been devoted to developing and understanding methods for training DNNs with less labeled data. We discuss the related approaches most relevant to this work below.

- Data augmentation. DA is commonly used to virtually expand the size of the labeled training set, making the trained network more robust to the variations introduced by the augmentation functions. Many segmentation networks rely on simple augmentation functions,

such as affine transforms, elastic transforms, flipping, and contrast jitter,[1,2,25,26] though more sophisticated methods for automating the augmentation process have been proposed.[27,28] In this work, we build on the uncertainty-based random sampling strategy proposed by Wu et al.[29] to select which augmentations to apply for each input during training.

- Weak supervision and PL. To learn from large datasets that do not have manual labels, weak supervision and PL approaches have been developed to automatically assign training labels. In weak supervision,[30,31] multiple programmatic labeling sources are developed to label each unlabeled data point in the training set; the outputs of the programmatic labeling sources are then aggregated into weak labels to train downstream models. In PL,[32,33] training labels for unlabeled data are generated from a network trained with a smaller set of labeled data. Each of these approaches generates a large, imperfectly labeled dataset for training downstream networks.
- Self- and semi-supervision. Self-supervised approaches learn from unlabeled datasets by constructing supervision tasks from the data itself. These self-supervised networks can then be fine-tuned for downstream tasks using small sets of labeled data. In image analysis, contrastive learning has achieved strong self-supervised performance while drastically reducing the required amount of labeled data.[34–38] Related approaches that aim to learn similar representations of the same image with different augmentations have also seen strong performance on a variety of downstream tasks.[39,40] Semi-supervised learning with a consistency loss (CL) is a similar approach that jointly trains on unlabeled and labeled data, encouraging the network to learn similar outputs for different augmentations of the same unlabeled image.[41–47] Finally, pseudolabels can be generated while enforcing that pseudolabels assigned to different augmentations of the same image are similar.[48]

We build on many of these ideas by combining DA, a CL, and PL into one pipeline for medical image segmentation.

2.3 *Reducing Labeled Data Needed for Medical Image Segmentation*

Some prior work has focused specifically on reducing labeled data requirements in the medical image segmentation setting. The most relevant to the work presented here includes an adaptation of SimCLR to medical image segmentation by pretraining a network backbone using a contrastive loss and proposing a new method to sample positive and negative image regions.[10] Weak supervision has also been adapted to medical image segmentation via a new graphical model that produces weak labels using multiple few-shot segmentation networks.[49] Other prior work has shown that noisy labels can be used to train downstream networks,[5,50] and past semi-supervised approaches use a CL over the segmentation masks predicted for an image with different augmentations applied.[6] Finally, a past framework has incorporated semi-supervised learning with unsupervised domain adaptation and noisy label learning.[51] We continue to build on these lines of work by combining DA, a CL, and PL into one framework while tracking uncertainty in each step and focusing on thorough evaluation of the semi-supervised networks with large clinical datasets over multiple segmentation targets and imaging views using clinical evaluation metrics.

3 Methods

We first describe our method for training segmentation networks with limited labeled data, then describe our segmentation targets and evaluation metrics. We then discuss the datasets we use in this work and end with details on our training procedure and data analysis.

3.1 *Description of Proposed Training Method*

Building on previous work in DA, consistency regularization, and PL, we propose a semi-supervised framework that uses unlabeled data and a small amount of labeled data to train medical image segmentation networks. This framework is built to be a flexible system that incorporates

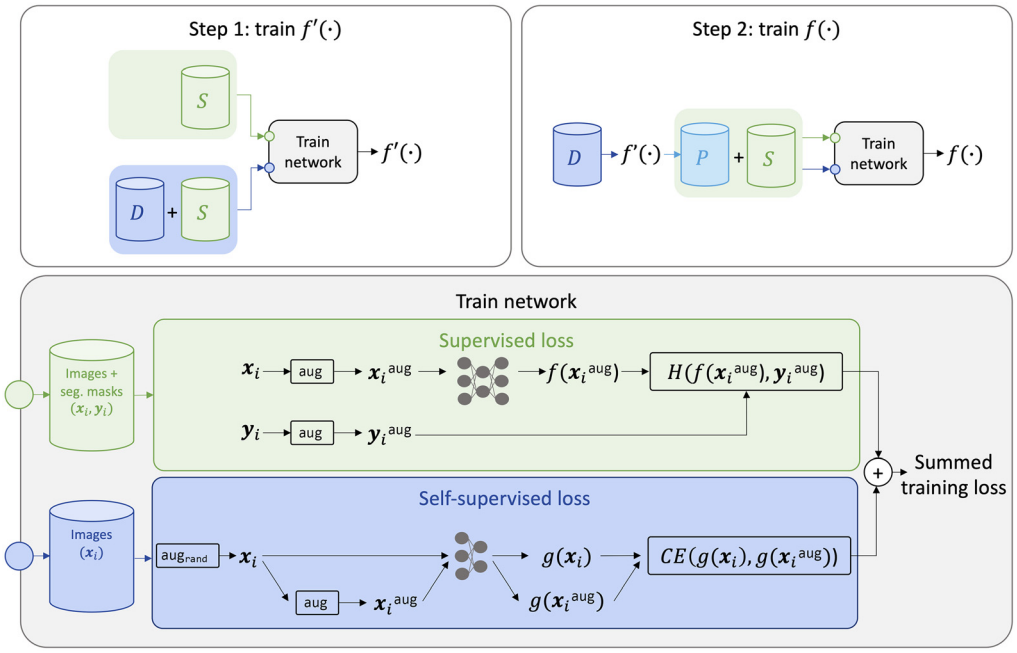

Fig. 1 Diagram of our proposed method. In step 1, we train an initial segmentation network $f'(\cdot)$ using a supervised loss over a small set of manually annotated data (S) summed with a self-supervised loss over all labeled (S) and unlabeled data (D). In step 2, we use $f'(\cdot)$ to generate a pseudolabeled dataset (P), comprising all unlabeled training images. We then train the final network $f(\cdot)$ using a supervised and self-supervised loss over the labeled and pseudolabeled data.

multiple sources of model supervision in a unified framework. Our proposed approach proceeds in two steps (Fig. 1).

- In step 1, we train an initial segmentation network $f'(\cdot)$. We sum a supervised cross-entropy loss—which learns from the small set of labeled data—with a self-supervised CL, which learns from the abundant unlabeled data. We select DAs for the supervised and self-supervised losses using an uncertainty-based random sampling approach.
- In step 2, we use $f'(\cdot)$ to generate pseudolabels for all unlabeled training data. We join the pseudolabels with the small set of manually annotated data, forming a larger labeled training set that we use to train the final segmentation network $f(\cdot)$.

We describe these steps in detail below.

We are given a set of labeled training data S and unlabeled training data D. The labeled training set consists of M corresponding images and segmentation masks, $S = (x_i, y_i)_{i\in[1,M]}$, where x_i is an image and y_i is its corresponding segmentation mask. The unlabeled training set consists of N images, $D = (x_i)_{i\in[1,N]}$, and $M < N$. We aim to learn a network $f(\cdot)$ to generate the segmentation mask given an input image.

Step 1: train pseudolabeler $f'(\cdot)$.

In the first step, we learn an initial segmentation network $f'(\cdot)$ by minimizing the sum of a supervised loss with DA over the labeled training data S and a self-supervised loss over the unlabeled training data D.

Using the labeled training data, we compute the pixel-wise cross-entropy loss $H(\cdot,\cdot)$ between the predicted segmentation probabilities output from $f'(\cdot)$ and the ground truth segmentation mask for each sample in S. We implement an uncertainty-based random sampling approach[29] to apply DA (augmentation procedure described fully in the Sec. 6.1.2).

We additionally incorporate a self-supervised loss to train $f'(\cdot)$ using unlabeled images. Specifically, we use a CL, which encourages model embeddings to be the same for perturbed

model inputs. We perturb the input by applying a segmentation-invariant augmentation, which we define as functions that do not change the segmentation mask (e.g., brightness adjustment, contrast jitter, histogram equalization), and compute the cosine embedding loss $\mathrm{CE}(\cdot,\cdot)$ between the original and perturbed model embeddings, which are taken from the penultimate layer of the model. We refer to the embeddings of $f'(\cdot)$ as $g'(\cdot)$. We extend the uncertainty-based random sampling approach to the self-supervised setting to select which augmentation to apply, described fully in the Sec. 6.1.3.

Summary of step 1. We train the pseudolabeler $f'(\cdot)$ by minimizing the sum of the supervised and self-supervised loss:

$$L' = \sum_{(\boldsymbol{x}_i,\boldsymbol{y}_i)\in S} H(f'(\boldsymbol{x}_i^{\mathrm{aug}}),\boldsymbol{y}_i^{\mathrm{aug}}) + \propto \sum_{\boldsymbol{x}_i\in S\cup D} \mathrm{CE}(g'(\boldsymbol{x}_i), g'(\boldsymbol{x}_i^{\mathrm{aug}})), \tag{1}$$

where $\propto$ is a weighting parameter. Each augmentation is chosen to maximize uncertainty.

Step 2: use pseudolabels to train the final network $f(\cdot)$.

In the second step, we begin by using $f'(\cdot)$ to predict a segmentation mask y_i' for each unlabeled image $x_i \in D$. We take the predicted segmentation masks as soft pseudolabels, forming the pseudolabeled set $P = (x_i, y_i')_{i\in[1,N]}$. Next, we join P and S to obtain a larger training dataset composed of both the pseudolabels and ground truth labels. We use this larger dataset to train the final segmentation neural network $f(\cdot)$, training on both the pseudolabels and ground truth labels simultaneously. We use the same training loss described in step 1, except that we modify the cross-entropy loss to preserve the uncertainty in the probabilistic pseudolabels output from $f'(\cdot)$ using the soft cross-entropy loss. Additionally, we ignore highly uncertain pseudolabels by ignoring pixels with pseudolabels between $0.5-\beta$ and $0.5+\beta$, where β is a hyperparameter we set to 0.05. Note that we do not initialize $f(\cdot)$ with the weights of $f'(\cdot)$.

Summary of step 2. We train the final segmentation neural network $f(\cdot)$ using the small set of manually annotated data and the larger set of pseudo labeled data by minimizing the following loss:

$$L = \sum_{(\boldsymbol{x}_i,\boldsymbol{y}_i)\in S} H(f(x_i^{\mathrm{aug}}), y_i^{\mathrm{aug}}) + \sum_{(\boldsymbol{x}_i,\boldsymbol{y}_i')\in P} H_{\mathrm{soft}}(f(x_i^{\mathrm{aug}}), y_i'^{\mathrm{aug}}) + \propto \sum_{\boldsymbol{x}_i\in S\cup P} \mathrm{CE}(g(x_i), g(x_i^{\mathrm{aug}})). \tag{2}$$

3.2 *Segmentation Targets and Evaluation Metrics*

We evaluate our methods on cine CMR, a widely used cardiac imaging procedure. Cine CMR acquires images of the heart throughout the cardiac cycle, allowing clinicians to evaluate cardiac function at different cardiac phases. The most common type of cine CMR is SAX imaging, which can capture multiple imaging planes at many time points. 2CH and 4CH LAX cine imaging is often also performed to view additional planes of the heart, where these LAX views image the heart at one imaging plane and many time points.

We segment the endocardium at the end diastolic (ED) and end systolic (ES) phases on SAX, 2CH LAX, and 4CH LAX and the epicardium at the ED phase on SAX. These various imaging views and segmentation targets enable us to evaluate our network over many biomarkers and demonstrate the flexibility of the approach by training networks over different structures and imaging views.

We analyze five cardiac functional biomarkers derived from the segmentation masks: the LV ejection fraction (EF), ED volume (EDV), ES volume (ESV), stroke volume (SV), and LVM. These are functional measures that quantitate the structure or function of the heart and are commonly used clinical metrics that inform patient care. Additionally, we report the Dice coefficient as a standard segmentation metric. Descriptions of and equations for each metric are included below.

- LV EDV measures the volume delineated by the LV endocardium segmentation at the ED phase.

- LV ESV measures the volume delineated by the LV endocardium segmentation at the ES phase.
- LV SV measures the volume of blood ejected from the LV during each contraction:

$$\mathrm{SV} = \mathrm{EDV} - \mathrm{ESV}. \tag{3}$$

- LV EF measures the fraction of blood ejected from the LV during each contraction:

$$\mathrm{EF} = \frac{\mathrm{SV}}{\mathrm{EDV}}. \tag{4}$$

- The LVM is the mass of the LV myocardium. We compute the LVM as the epicardium delineated volume minus the endocardium delineated volume at the ED phase, multiplied by an estimate of the myocardial density (1.05 $\mathrm{g/mm^3}$).
- Dice coefficient measures the overlap between two segmentation masks. At minimum, Dice takes the value 0 (achieved when there is no overlap between the two masks); at maximum Dice takes the value 1 (achieved when the two masks are identical). The Dice coefficient between two binary segmentation masks, y_i and y_j, is defined as:

$$\mathrm{Dice}(y_i, y_j) = \frac{2|y_i \cap y_j|}{|y_i| + |y_j|}. \tag{5}$$

3.3 *Dataset Descriptions*

We use three previously curated datasets in this retrospective study.

To train the networks and compare against segmentation masks drawn by an expert, we use the contoured dataset. This dataset contains SAX and LAX cines collected at multiple healthcare centers and includes healthy and diseased patients, multiple magnet strengths, and multiple scanner manufacturers.[52] The standard balanced steady state free precession (BSSFP) 2D cine sequences were used to acquire these images, with the standard imaging parameters: BSSFP, FOV 360×270 $\mathrm{mm^2}$, TR 2.7 ms, acceleration $R = 2$, flip angle 50 deg. Each cine slice was individually placed and acquired separately. Each image in this dataset has a segmentation mask drawn by an expert which delineates the endocardium on the ED and ES SAX and LAX images and the epicardium on the ED SAX images. We split this dataset randomly by patient into 60% training, 20% validation, and 20% testing. We report the total number of scans, phases, and 2D image slices for each imaging view and segmentation target in Table 1.

We use an additional two datasets for evaluation. First, we evaluate on a scan-rescan dataset, which contains 108 patients collected from multiple healthcare centers, scanners, field strengths, and represents patients with different diseases (myocardial infarction, left ventricular hypertrophy, cardiomyopathy, other pathologies, and healthy volunteers).[4] Each patient in this dataset has five clinical biomarkers measured by an expert cardiologist (15+ years of experience). Additionally, each patient was scanned twice to evaluate the scan-rescan variation (median scan interval 1 day, 82% scanned on the same day). We use this dataset to compare against the expert's computation of the five clinical biomarkers and to compute the within-subject CV from each patient's repeated scans.

Finally, we evaluate on a generalization dataset, which contains 1223 patients with SAX and LAX cines collected from University of Pittsburgh Medical Center.[53] Each patient has clinical biomarkers measured by an expert. We use this dataset to assess the ability of networks to generalize to an external healthcare center, as no images in the contoured dataset (i.e., the training dataset) were collected at the University of Pittsburg Medical Center.

The scan-rescan dataset is publicly available in Ref. 54. The contoured and generalization datasets are not currently available publicly, though additional information about the data collection, characterization, and access can be found in the original studies.[4,52,53] All datasets were acquired with the required ethical or secondary audit use approvals or guidelines (as per each center) that permitted retrospective analysis of deidentified data for technical development, protocol optimization, and quality control. Written consent had been obtained

Table 1 Summary of data in the contoured dataset.

	SAX endocardium	SAX epicardium	2CH LAX endocardium	4CH LAX endocardium
Number of scans	1208	1208	622	622
Number of labeled phases per scan	2	1	2	2
Number of 2D slices per phase (average)	11.61	11.61	1	1
Total number of 2D slices in the dataset	28050	14025	1244	1244
Total number of 2D slices in the train split	16812	8406	746	746
Total number of 2D slices in the val split	5540	2770	248	248
Total number of 2D slices in the test split	5698	2849	250	250

The contoured dataset contains labeled data for each segmentation task. The number of scans, phases, and imaging planes for each task are reported here, which are multiplied together to compute the total number of 2D image slices in each dataset. Each dataset is split randomly by patient into 60% train, 20% validation, and 20% test.

for all subjects and study was approved by the institutional review committee of each center.

3.4 *Training Details*

We use a modified U-Net architecture previously developed for CMR segmentation and available on an open-source platform.[55–57] This architecture is a 2D segmentation model. For the SAX segmentation tasks, which contain multiple imaging planes per image, we train and predict on each SAX cine's 2D image slices independently; the total number of 2D image slices in each split of the contoured dataset is reported in Table 1.

To evaluate the semi-supervised learning method, we compare three training approaches.

- Proposed-100 models. Using our proposed semi-supervised approach, we train segmentation models using 100 labeled 2D image slices and the remainder of the training set as unlabeled images. We will refer to these as the proposed-100 models.
- Naïve-100 models. For comparison, we use the same 100 labeled image slices and no semi-supervision to train the naïve-100 models. These models were trained with simple DA (i.e., for each image during training, one of all possible DA operations is randomly selected and applied) and do not leverage any unlabeled data.
- Fully supervised models. We also train fully supervised segmentation networks, which are trained using all labeled data in the contoured training split. These models represent the expected upper bound of performance.

For models that do not use all of the labeled training data (e.g., proposed-100 and naïve-100), we randomly select which training scans are used as labeled vs. unlabeled data. For multi-slice data (e.g., SAX), we select one slice from each patient's cine, ensuring that basal, middle, and apical slices are equally represented. In the Sec. 6.2.2, we explore an alternative sampling strategy. For datasets with both ED and ES phases, we similarly ensure that ED and ES phases are equally represented. Hyperparameters and additional training details are listed in Sec. 6.1.4.

3.5 *Data Analysis*

All code and analyses were written in Python 3.6.1. We report the mean Dice coefficient with 95% CIs computed via bootstrapping using the seaborn package (v0.11.1).[58] Dice scores measure the overlap between two segmentation masks, with a value of 1 indicating perfect alignment. We use Lin's CCC, an intraclass correlation coefficient, to evaluate the

inter-observer agreement between the expert and models for each biomarker. Lin's CCC captures the agreement between two sets of measurements as well as the deviation from the line of perfect concordance.[59] Higher Lin's CCC indicate greater agreement, with values near 1 indicating strong agreement, values near −1 indicating strong discordance, and values near 0 indicating no concordance. Additionally, we report the precision of the expert and model on the scan-rescan dataset by computing the within-subject CV using the root mean square method.[60] The within-subject CV is used to measure the variation of repeated measures on the same subject. Here, lower coefficients of variation are desired, as we want lower variation when using the same method to repeatedly measure the same subject. We report 95% confidence intervals (CIs) of the within-subject CV, computed by bootstrapping (1000 iterations). Finally, we run a repeated-measures ANOVA test to assess the effect of the training method on the Dice score. We run this test using the statsmodels Python package (v0.12.2).[61]

4 Results

4.1 *Overview of Semi-Supervised Model Performance*

While we focus the majority of our evaluation on clinical metrics, we first report the mean Dice coefficients of the proposed-100 and fully supervised networks in Table 2 as standard segmentation metrics. Histograms of the Dice coefficients are provided in Fig. 6. The proposed-100 models achieve within 0.73% and 0.53% of the fully supervised performance for the SAX epicardium and 4CH LAX endocardium tasks and exceed the mean performance of the fully supervised 2CH LAX endocardium task.

Figure 2 and Fig. 7 show the proposed-100 segmentation masks that achieve median and high Dice coefficients, respectively. Qualitatively, we see strong agreement between the ground truth and predicted segmentation masks. Additionally, we compare against previously published self- and semi-supervised training methods in the Sec. 6.2.4.

4.2 *Evaluation of Cardiac Functional Parameters*

We next evaluate these proposed-100 models on the scan-rescan and generalization datasets to assess semi-supervised network performance on clinically used metrics.

4.2.1 *Scan-rescan evaluation*

We compare five cardiac functional measures computed by the proposed-100 networks to an expert cardiologist. Figures 3(a) and 3(b) show the correlation and Bland-Altman plots for each biomarker (the mean and standard deviations of these values are reported in Table 6). To assess the agreement between the predicted and expert biomarkers, we use Lin's CCC and find strong agreement for all metrics (Table 3). Additionally, we report Lin's CCC for the naïve-100 models,

Table 2 CMR segmentation performance.

Dice coefficient, contoured dataset	SAX endocardium	SAX epicardium	2CH LAX endocardium	4CH LAX endocardium
Proposed-100	0.897 (0.072)	0.948 (0.021)	0.950 (0.038)	0.941 (0.055)
Fully supervised	0.907 (0.069)	0.955 (0.021)	0.948 (0.040)	0.946 (0.045)
Labeled training data reduction	99.4%	98.8%	86.6%	86.6%

Mean Dice coefficients (standard deviation) over the test set of the contoured dataset for each of the four segmentation targets: SAX endocardium, SAX epicardium, 2CH LAX endocardium, and 4CH LAX endocardium. We also report the training data labeling reduction achieved by proposed-100 compared to the corresponding fully supervised network.

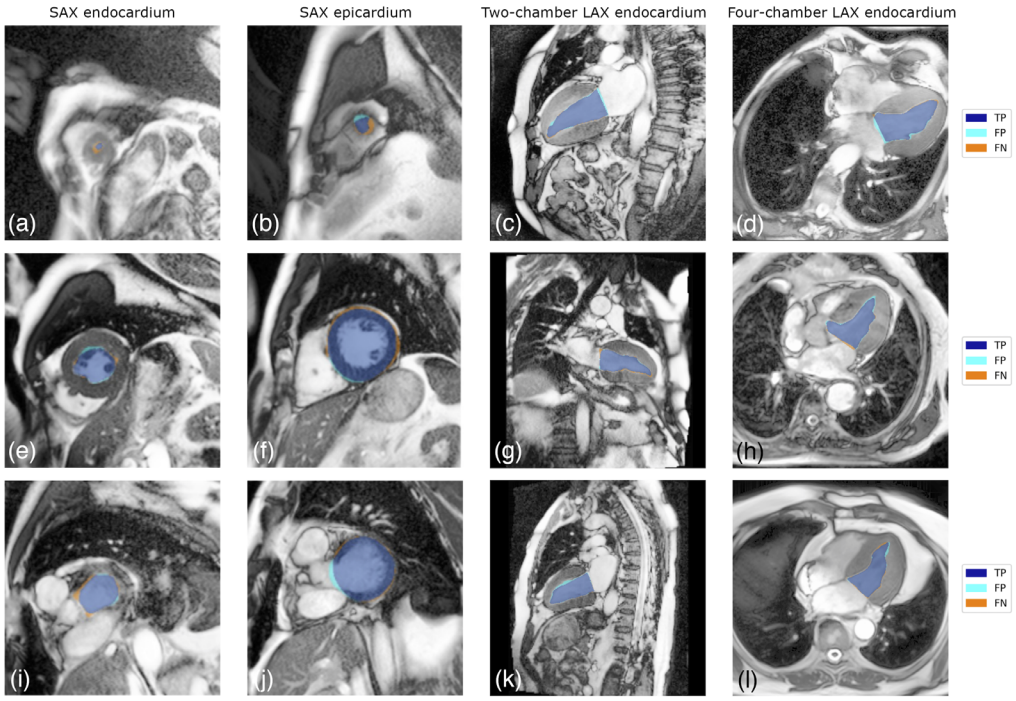

Fig. 2 Visualization of segmentation masks predicted by the proposed-100 models for the four segmentation targets; these segmentation masks achieve the median Dice coefficients in the test sets. The predicted true positives (TPs), false positives (FPs), and false negatives (FNs) are shown for each image. For the SAX segmentation targets, apical (a, b), middle (e, f), and basal (i, j) slices from the same patient are shown. For the LAX segmentation targets, three different patients are shown.

which were trained with the same 100 labeled images as proposed-100 but without the proposed semi-supervised learning method.

For context, we also include the correlation and Bland-Altman plots of the intra-observer and inter-observer variation for this dataset in Fig. 8 (as collected in prior work[4]); Lin's CCC for the inter-observer variation is 0.91, 0.98, 0.98, 0.92, and 0.94 for EF, EDV, ESV, SV, and LVM.

Next, we report the scan-rescan variation of the biomarkers computed by the proposed-100 networks, which measures the precision: for a patient scanned twice in close succession, how reproducible are the biomarkers given the same annotation method. Reproducibility is clinically important because these biomarkers are used to assess treatment efficacy and disease progression.[4] We assess the scan-rescan variation by comparing the biomarkers from each patient's two scans using the within-subject CV. In Table 4, we report the within-subject CV and the 95% CIs for the models and the expert. We see that for all metrics, the naïve model has higher CVs than the proposed-100 model and the expert annotator. In contrast, the proposed-100 model has better average performance than the expert annotator on the ESV and EF metrics and is within the 95% CI of the expert for EF, EDV, ESV, and SV metrics.

4.2.2 *Generalization evaluation*

To assess how the proposed-100 networks generalize to data collected from an independent center that did not provide any training data, we use the proposed-100 networks to segment 1,223 SAX CMR cines collected at the University of Pittsburgh Medical Center. This was an important evaluation because past work has shown that models trained on small or homogenous datasets often experience degraded performance when deployed on real-world data.[62–67] Since semi-supervised networks use a small amount of labeled training data but a

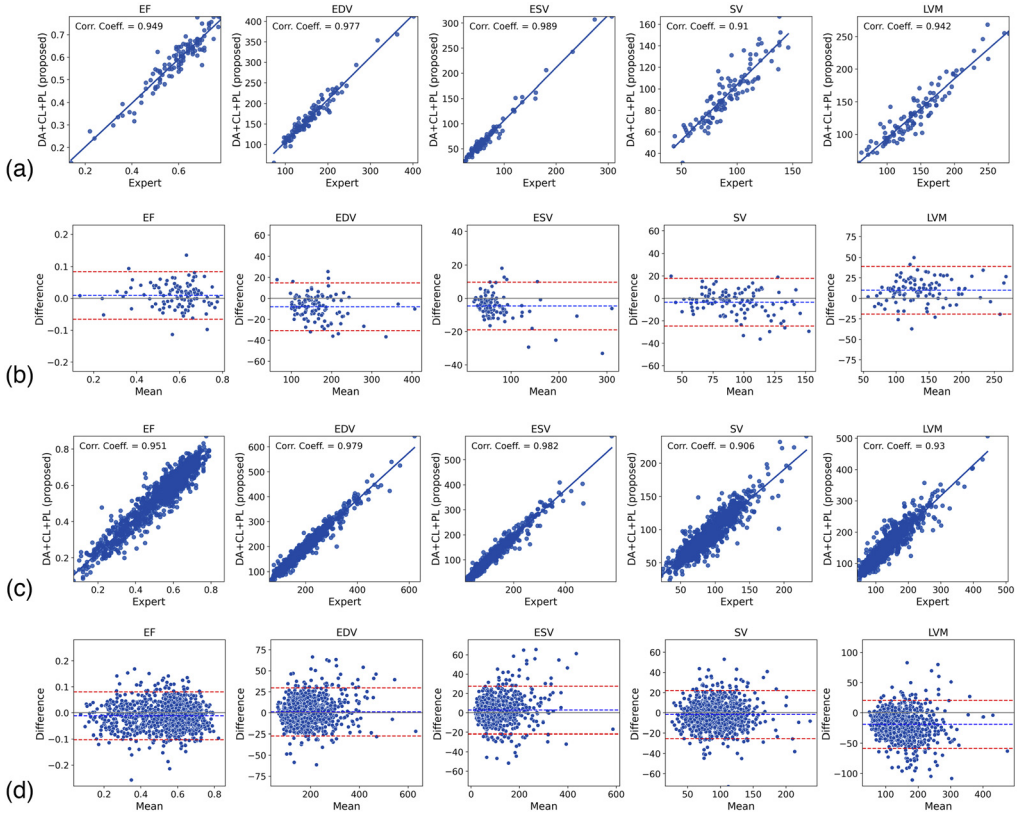

Fig. 3 Correlation and Bland-Altman plots showing the five cardiac imaging biomarkers. We compare the biomarkers computed by proposed-100 to the biomarkers computed by an expert on the scan-rescan dataset (a, b) and the generalization dataset (c, d).

Table 3 Lin's CCC on the scan-rescan dataset.

CCC, scan-rescan dataset	EF	EDV	ESV	SV	LVM
Naïve-100	0.89 (0.82 to 0.93)	0.91 (0.83 to 0.95)	0.96 (0.9 to 0.98)	0.81 (0.73 to 0.87)	0.80 (0.69 to 0.87)
Proposed-100	0.95 (0.91 to 0.97)	0.96 (0.94 to 0.98)	0.98 (0.96 to 0.99)	0.89 (0.85 to 0.92)	0.92 (0.88 to 0.94)

Lin's CCC (95% CIs) for the naïve training ("naïve-100") and the semi-supervised method ("proposed-100") for each imaging biomarker.

Table 4 Within-subject CV on the scan-rescan dataset.

CV, scan-rescan dataset	EF	EDV	ESV	SV	LVM
Naïve-100	8.44% (6.63 to 10.27)	7.04% (5.84 to 8.21)	11.41% (8.58 to 14.31)	12.33% (10.23 to 14.07)	8.06% (6.63 to 9.63)
Proposed-100	5.72% (4.80 to 6.63)	6.21% (5.01 to 7.40)	8.63% (7.42 to 9.73)	9.67% (8.00 to 11.42)	6.10% (4.95 to 7.17)
Expert	6.15% (5.24 to 7.05)	5.75% (4.76 to 6.70)	10.04% (8.27 to 11.86)	9.46% (7.80 to 10.86)	4.84% (4.13 to 5.59)

The within-subject CV (95% CI) for the naïve training ("naïve-100"), semi-supervised method ("proposed-100") and the expert annotator for each imaging biomarker.

Table 5 Lin's CCC on the generalization dataset.

CCC, generalization dataset	EF	EDV	ESV	SV	LVM
Naïve-100	0.90 (0.88 to 0.92)	0.95 (0.92 to 0.97)	0.96 (0.93 to 0.97)	0.87 (0.85 to 0.89)	0.80 (0.77 to 0.82)
Proposed-100	0.95 (0.94 to 0.95)	0.98 (0.97 to 0.98)	0.98 (0.97 to 0.98)	0.90 (0.89 to 0.92)	0.87 (0.85 to 0.89)

Lin's CCC (95% CI) for the naïve training ("naïve-100") and the semi-supervised method ("proposed-100") for each imaging biomarker on the generalization dataset.

large amount of unlabeled training data, it is not obvious how well they would generalize to new datasets.

Figures 3(c) and 3(d) show the correlation and Bland-Altman plots for each biomarker predicted by the proposed-100 models compared to an expert on the generalization dataset. We find high correlation coefficients and Lin's CCCs (Table 5) between the model and manual measurements. For comparison, we also evaluate the generalization performance of the naïve-100 models. This analysis shows that the naïve models achieve worse generalization performance than the semi-supervised models, with the naïve model's mean CCC falling outside the 95% CIs for all metrics compared to the semi-supervised models.

4.3 *Error Analysis*

Next, we discuss the three primary error modes we observed in the semi-supervised networks, illustrated in Fig. 4. We also compare these error modes against the fully supervised networks. Though the semi-supervised and fully supervised models achieve comparable Dice coefficients (Table 2), whether the semi-supervised networks experience different error modes than the fully supervised networks is an important consideration for model users.

The first error mode we notice is on apical and basal slices. Correctly segmenting the apical and basal slices in SAX CMR is a known challenge.[16] While the proposed-100 model achieves an aggregate 0.897 Dice for the SAX endocardium, the network achieves a mean 0.796 Dice on basal slices and drops to a Dice of 0.487 on apical slices. We observe that while the fully supervised network achieves a similar score on the basal slices (Dice of 0.808), it performs better on the apical slices (Dice of 0.701).

The second error mode we observe is on anatomical abnormalities. For example, on SAX and LAX images, we observe the lowest Dice coefficients often occur with abnormally small LV cavities, such as in hypertrophic cardiomyopathy (HCM, examples in Fig. 10). Quantitatively, we compare the Dice coefficient for patients whose ESV is in the highest and lowest 10% of our test dataset (Table 7) and observe up to a 7.36% performance drop. Looking at these same subsets with the fully supervised networks, we observe a similar pattern: the fully supervised network drops 7.79% for patients with small LV cavities.

Finally, we observe errors on images with artifacts or acquisition problems. Incorrect plane planning appears primarily in 4CH LAX, where the imaging plane misses the apex, and we observe both the proposed-100 models and the fully supervised models have low scores on these cases.

4.4 *Ablation Studies*

Finally, we perform ablations to understand how different amounts of training data and different sources of network supervision impact network performance. Specifically, we assess four different training methods (described below) where each method provides an additional layer of supervision to the network. This experiment allows us to assess how each component of our method contributes to the final performance.

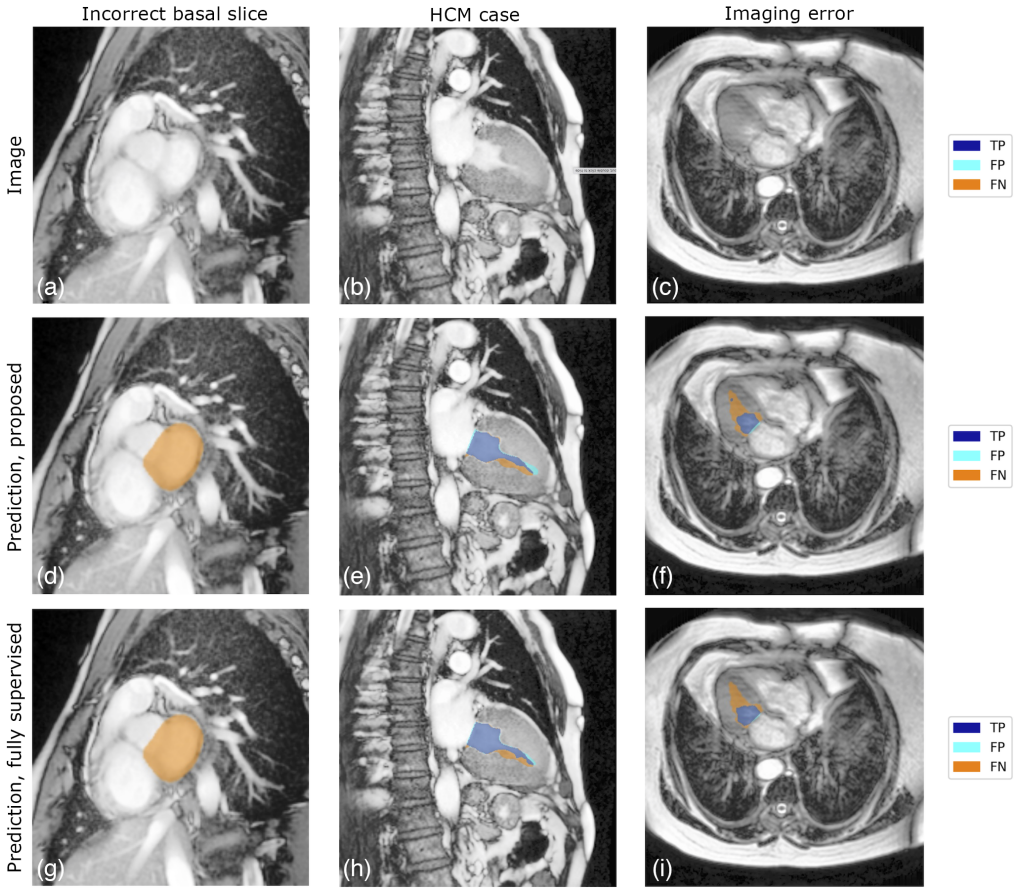

Fig. 4 In each column, we show an error mode observed in the predicted segmentation masks: apical/basal slices, patients with certain pathologies (e.g., hypertrophic cardiomyopathy, HCM), imaging artifacts/errors. In panels (a)–(c), we show the image alone; in panels (d)–(f), we show the segmentation mask predicted by our proposed network; and in panels (g)–(i), we show the segmentation mask predicted by the fully supervised network. The predicted TPs, FPs, and FNs are shown for each image.

- Naïve. This ablation represents the lower bound of expected performance as we use only the labeled data to train the network via the cross-entropy loss. We perform naïve DA (i.e., for each image during training, one of all possible DA operations is randomly selected and applied) and do not leverage any unlabeled data.
- "DA." This ablation uses only labeled data with a cross-entropy loss but improves the DA strategy by implementing the approach proposed by Wu et al.[29] and described fully in the Appendix. Briefly, we select which augmentation to perform for each image based on which augmented image the network is least certain about.
- "DA+CL." We build on the uncertainty-based random sampling DA scheme by adding a CL over all unlabeled data. This network represents the output of Step 1 in our method and is trained with the loss function given in Eq. (1).
- "DA+CL+PL." This is our proposed method, resulting in the network described in Step 2 above. Using the DA+CL network, we predict pseudolabels for all unlabeled data in our dataset. Then, we retrain the network by including the soft pseudolabels in the labeled dataset.

Additionally, we sweep the number of labeled training images supplied to each method. By sweeping the number of labeled images, we can evaluate how each training approach performs with varying amounts of labeled data. All ablations are trained with the same architecture and hyperparameters.

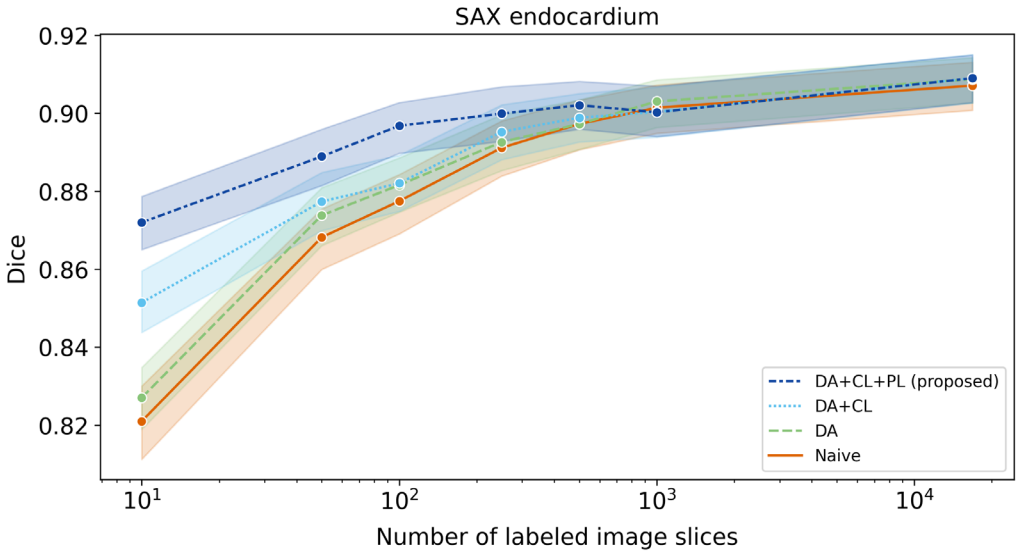

Fig. 5 Dice coefficient with 95% CIs for the SAX endocardium segmentation task. We train our proposed method using DA, a CL, and PL. We also show the performance of three ablations. We sweep the number of labeled training images from ten to 16,812 to show how performance changes as a function of labeled training images.

In Fig. 5, we show the performance of the proposed approach and each ablation on the SAX endocardium segmentation task as a function of number of labeled training images. We evaluate segmentation performance on the held-out test set using the Dice coefficient. We conduct a one-way repeated-measures ANOVA test to examine the effect of the training method on the Dice score. We repeat this test for each number of labeled training data represented in Fig. 5, including $n = 10$, 50, 100, 250, 500, 1000, and 16,812 labeled training images. We apply the Bonferroni Correction to account for the multiple statistical tests and set our p-value significance cutoff to 0.007 (=0.05/7). Below, we report the F statistic and p value for this test, where there are three degrees of freedom for the training method and 1395 deg of freedom for the error. Our results showed that the training method led to statistically significant differences in Dice scores for: $n = 10$ ($F(31,395) = 228.4600$, $p < 0.001$); $n = 50$ ($F(31,395) = 54.9570$, $p < 0.001$); $n = 100$ ($F(31,395) = 53.6677$, $p < 0.001$); $n = 250$ ($F(31,395) = 13.5300$, $p < 0.001$); and $n = 500$ ($F(31,395) = 6.2839$, $p < 0.001$) and statistically insignificant differences in Dice scores for: $n = 1000$ ($F(31,395) = 1.7491$, $p = 0.1551$) and $n = 16{,}812$ ($F(31,395) = 1.1068$, $p = 0.3452$).

5 Discussion

Although neural networks can automate segmentation, large and diverse training datasets are costly to label. In contrast, unlabeled data is often readily available. While recent work has utilized unlabeled data to train segmentation networks, few have focused on evaluating these methods on diverse medical imaging datasets using clinical biomarkers. In this work, we propose a semi-supervised framework that leverages uncertainty-based DA, a CL, and PL, and we focus on thoroughly evaluating these models on cardiac MR cine segmentation tasks using five functional biomarkers and multi-institutional datasets. We further discuss the error modes of the networks and characterize network performance with different amounts of labeled data and different sources of model supervision.

Through this evaluation, we observed multiple interesting takeaways. First, in our evaluation of the scan-rescan dataset, we observed strong performance on clinical metrics using the proposed-100 models: the semi-supervised models achieved high Lin's CCC and overlapping 95% CIs of the within-subject coefficients of variation compared to an expert annotator for four of the five functional biomarkers (Tables 3 and 4). In contrast, we found that the naïve models had

higher CVs and lower CCCs than the semi-supervised models for all clinical metrics we evaluate, with the CVs of the naïve models falling outside the 95% CIs of the expert annotator and the CCCs of the naïve models falling outside the 95% CIs of the semi-supervised models. This shows the impact of semi-supervision on the accuracy of the clinical metrics. Additionally, we observed that the semi-supervised pipeline improved generalization performance on data collected from an independent hospital compared to the naively trained models (Table 5). This strong performance by the semi-supervised networks is promising, as they were trained with only 100 labeled image slices (equivalent to approximately 10 patients' labeled data for the volumetric scans). Such low labeling requirements make training segmentation networks much more accessible than relying on very large labeled datasets, particularly considering the burdensome process of labeling segmentation data.

We further observed that combining multiple sources of supervision in the pipeline led to the highest network performance (Fig. 5); for example, adding the CL over all unlabeled data improves performance at the most limited labeled data setting (i.e., 10 labeled images) by 3.65% compared to the naïve method and further adding pseudo labels improves performance by 6.21% over the naïve method. The PL provided the greatest impact on network performance, which is a simple training step that can be appended to other training pipelines as well. We observe the performance gains beyond ~250 labeled images with the full semi-supervised pipeline were minimal. While we expect the scale of the x-axis to change for different segmentation targets, this scaling analysis informs how developers should use the training pipeline and quantifies the value of additional labeling effort when using the semi-supervised pipeline. Labeling a few training images results in the largest gains and labeling reduction. Users who need additional performance can label additional training points, though the relative performance improvements will be less. Our results also suggest that for users with large, labeled datasets, semi-supervision may not useful. However, these results should be interpreted with the knowledge that even a modest increase in Dice coefficient can be important and lead to higher accuracy on clinical metrics (as shown in the difference in performance on clinical metrics between the naïve and proposed models).

However, we did observe an error mode exaggerated by semi-supervised training. Although the semi-supervised and fully supervised SAX endocardium networks achieved similar Dice coefficients, the models' Dice scores on apical slices differed significantly. This error mode may be important for some downstream applications, such as monitoring apical HCM. This result suggests that targeted evaluation and additional labeling may be required for smaller structures (e.g., apical slices) when using limited-label training methods, while larger structures (e.g., left ventricular cavity) can be segmented using few labeled examples, achieving similar performance to fully supervised systems even on pathological cases (e.g., small left ventricular cavities). This error mode in combination with our observations from the scaling analysis also suggest an interesting direction for future work: by combining semi-supervised training methods with human-in-the-loop iterative labeling, the labeling of "simpler" structures that are easier for networks to learn (e.g., mid slices in SAX cines) can be automated earlier in the data curation process with very few labeled images, enabling the human expert to spend more time labeling difficult structures (e.g., apical slices).

5.1 *Limitations*

This work relies on randomly sampled labeled data; more purposefully selecting training images to label may lead to higher performance. Additionally, while we are reducing the size of the labeled training set, we still rely on a manually labeled validation set. Compared to purely self-supervised methods, our method is more computationally expensive to retrain with new labeled data or for a different task: since we rely on labeled data in both the first and second step of our method, adding labeled data requires retraining the entire system instead of only repeating the final fine-tuning step as is typical in self-supervised systems. Additionally, self-supervised pretraining has the advantage of being task-agnostic. However, in settings with a defined segmentation task and limited labeled data, the semi-supervised method outperforms existing self-supervised methods. Finally, we only perform retrospective evaluation in this study.

5.2 *Conclusion*

In this study, we focused on evaluating semi-supervised networks for medical image segmentation using large, heterogenous datasets and clinical metrics. As methods for training models with little labeled data become more common, we believe this knowledge about how they perform on clinical tasks, how they may fail, and how they perform with different amounts of labeled data is useful to model developers and model users. While this study focuses on cardiac cine segmentation tasks, the semi-supervised training pipeline can be directly applied to other organs and different imaging modalities.

6 Appendix

6.1 *Methods*

6.1.1 *Image preprocessing*

All CMR images were resampled to 1 mm^2. The SAX images were preprocessed with an N4 bias field correction using the Simple ITK package in Python,[68] then cropped to 224×224 around the image center. The LAX images were cropped to 416×416. Each 2D image slice was preprocessed with a histogram equalization operation.

6.1.2 *Uncertainty-based data augmentation: supervised loss*

We use DA to virtually expand the size of the labeled training set. We want to apply an augmentation function to each image and segmentation mask (x, y) to obtain the augmented pair $(x^{\text{aug}}, y^{\text{aug}})$. To select which augmentation to apply to each sample, we implement the uncertainty-based random sampling approach presented in previous work.[29] Using this augmentation method, we first define K possible augmentation transformations (listed in the Sec. 6.1.4). For each image x, we randomly select C of those transformations. Each transformation is applied separately to (x, y), resulting in C augmented pairs, $\{(x_c^{\text{aug}}, y_c^{\text{aug}})\}_{c=1}^{C}$. The network $f'(\cdot)$ predicts the segmentation mask for each augmented training sample. The augmentation that results in the greatest uncertainty (i.e., highest cross-entropy loss) is chosen as the augmentation to apply for that (x, y) pair. This augmentation approach provides a principled method for choosing which augmentation function to apply to each input image; intuitively, this approach forces the network to improve augmentation examples that it is the least certain about.

6.1.3 *Uncertainty-based data augmentation: self-supervised loss*

To apply consistency regularization, we need to select the segmentation-invariant augmentation to form the pair (x, x^{aug}). Extending previous work in the supervised setting,[29] we implement another uncertainty-based random sampling approach to select augmentations in the self-supervised setting. For each image x, we first randomly apply one of the K possible augmentation transformations we used in the supervised setting. Then, we define a set of L possible segmentation-invariant augmentation transformations. We randomly select C of these segmentation-invariant transformations and apply each transformation separately to x, resulting in the pairs $\{(x, x_c^{\text{aug}})\}_{c=1}^{C}$. Let $g'(\cdot)$ be a truncated $f'(\cdot)$; in this work, we remove the final convolutional layer of $f'(\cdot)$ to get $g'(\cdot)$. We compute the cosine embedding loss $CE(\cdot, \cdot)$ for each pair of embeddings $\{(g'(x), g'(x_c^{\text{aug}}))\}_{c=1}^{C}$, which assesses similarity of the embeddings. The segmentation-invariant augmentation that results in the greatest uncertainty (i.e., greatest cosine embedding loss) is chosen as the augmentation used to form the positive pair (x, x^{aug}).

6.1.4 *Additional training details*

The set of augmentations we select from for the supervised loss include: maximizing image contrast, enhancing image contrast, brightness jitter, contrast jitter, elastic transform, cutout,

equalize, invert, rotate, sharpness jitter, shear, translate, horizontal flip, vertical flip, and affine transform. The set of augmentations we select from for the self-supervised loss include: maximizing image contrast, brightness jitter, enhancing image contrast, contrast jitter, and equalize.

To select hyperparameters, we trained a fully supervised SAX endocardium segmentation network with an Adam optimizer using a grid search over the following hyperparameters: learning rate of $1e^{-3}$, $1e^{-4}$, and $1e^{-5}$; weight decay of 0, $1e^{-3}$, and $1e^{-5}$; learning rate scheduler of none and linear decay. We selected a learning rate of $1e^{-4}$ and a weight decay of $1e^{-5}$. The learning rate was set to decay using linear learning rate decay with a minimum learning rate of $1e^{-7}$. The weighting parameter balancing the supervised and self-supervised loss was tuned separately among 1, 3, and 5, then set to 1. Batch size was set to 4. The number of considered augmentations C was set to 4.

6.2 *Results*

6.2.1 *Supporting data*

In this section, we include supporting data for results reported in the main text.

In Fig. 6, we plot the histograms of Dice coefficients for each task trained with the proposed-100 models. In Fig. 7, we plot the images and predicted segmentation masks that achieve the highest Dice coefficients in the test set. In Fig. 8, we show the correlation and Bland-Altman plots for inter-observer (top two rows) and intra-observer (bottom two rows) variation on the scan-rescan dataset. In Fig. 9, we plot the difference in Dice coefficients achieved by the fully supervised networks and the proposed-100 models. Finally, in Fig. 10, we show examples of patients with small left ventricular cavities, which the models struggle to correctly segment.

In Table 6, we provide the mean cardiac functional parameters computed by the model, a human trainee, and an expert. In Table 7, we compare the Dice coefficients for models on patients with high vs. low ESVs, where we notice model performance drops on patients with

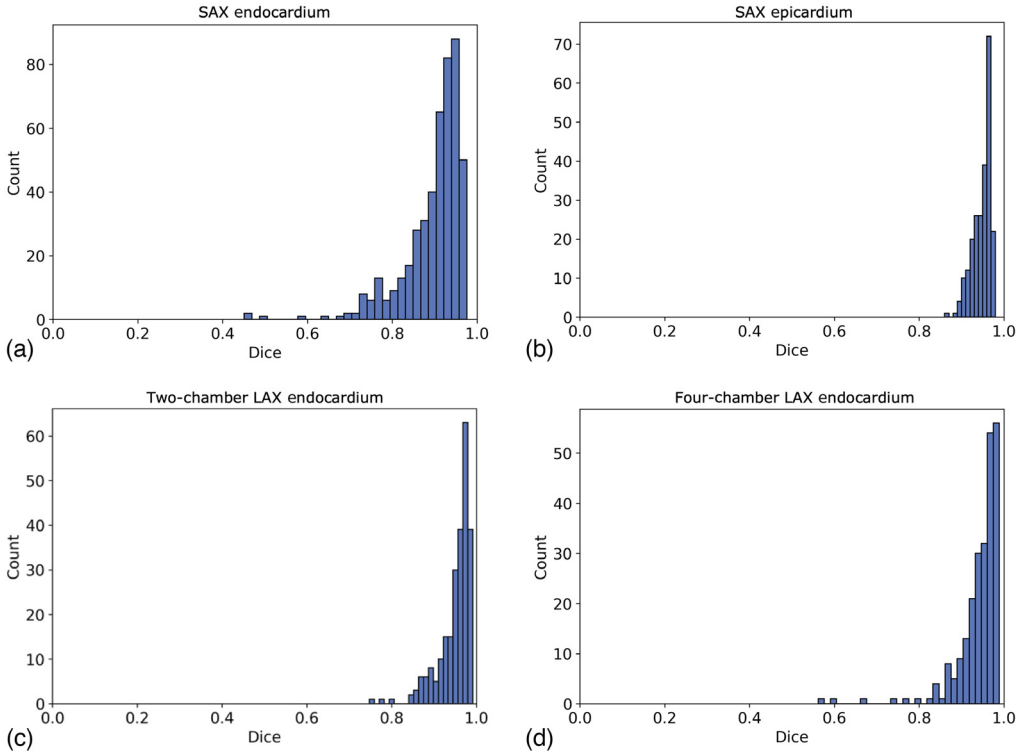

Fig. 6 Histogram of Dice coefficients for the proposed-100 networks on the test set for each segmentation target: (a) SAX endocardium, (b) SAX epicardium, (c) 2CH LAX endocardium, and (d) 4CH LAX endocardium.

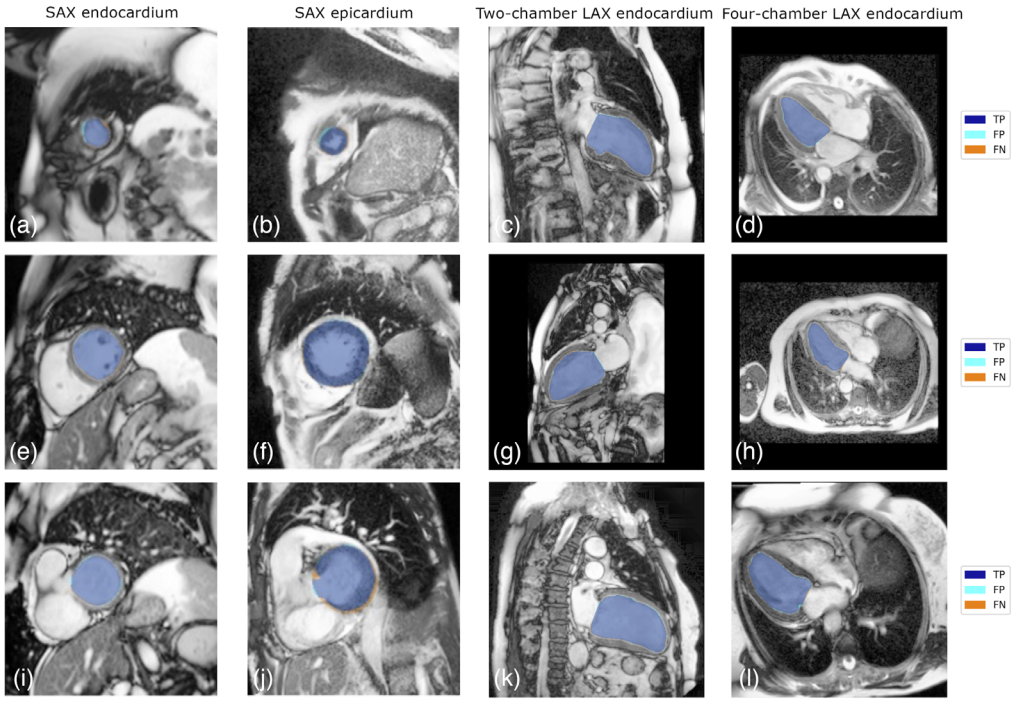

Fig. 7 Visualization of segmentation masks predicted by the proposed-100 models for the four segmentation targets; these segmentation masks achieve some of the highest Dice coefficients in the test sets. The predicted TP, FP, and FN are shown for each image. For the SAX segmentation targets, apical (a, b), middle (e, f), and basal (i, j) slices from the same patient are shown. Three different patients are shown for the 2CH LAX segmentation targets (c, g, k) and 4CH LAX segmentation targets (d, h, l).

low ESVs. Finally, in Table 8, we provide the mean Dice coefficients achieved by each baseline for different amounts of training data, as plotted in Fig. 5.

6.2.2 *Sampling method*

In the main text, we randomly selected which patients we used as the labeled data set. For volumetric images (e.g., SAX cine), we selected a single 2D slice from each randomly selected patient. For example, when training networks with 100 labeled image slices, we randomly selected 100 patients with SAX cines. We then chose a single 2D image slice from each patient's volumetric SAX. This sampling strategy provides the network with greater anatomical variation in the labeled dataset. To examine the sampling strategy's impact on performance, we include results here from an alternate sampling strategy. Instead of selecting single 2D image slices from each patient, we include all image slices in the patient's SAX cine. For example, to train a network with ∼100 labeled image slices, we randomly selected 9 patients with SAX cines. We then include all of the 2D image slices from each patient's volumetric SAX cine in the labeled training set (totaling 98 2D image slices among the 9 randomly selected patients). The results for the two sampling strategies are in Table 9. These results show the sampling strategies we assessed have minimal impact on the mean Dice coefficient.

6.2.3 *Uncertainty-based random sampling approach to data augmentation*

As described in the Methods section, we use an uncertainty-based random sampling approach to select the augmentations to apply to each input image for both the supervised cross-entropy loss and the self-supervised CL. In Fig. 5, we show the impact of the uncertainty-based random sampling augmentation approach on the supervised loss by plotting the naïve method (without uncertainty-based random sampling) against the DA method (with uncertainty-based random

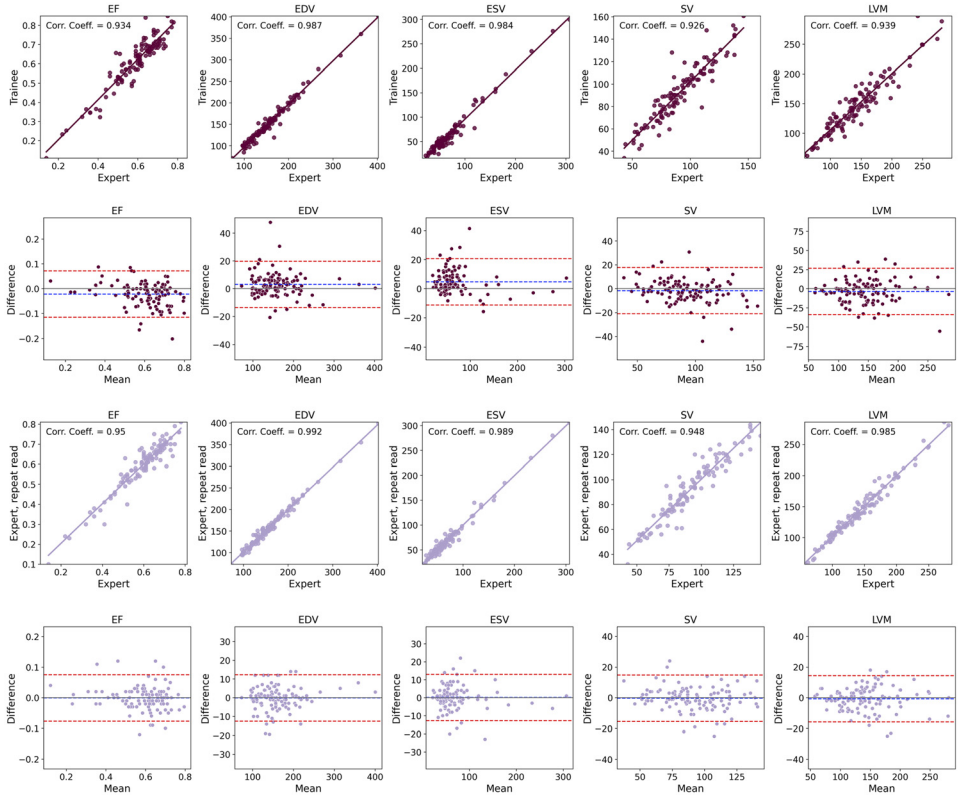

Fig. 8 Correlation and Bland-Altman plots showing the five cardiac imaging biomarkers. Top two rows: we compare the biomarkers computed by an expert to the biomarkers computed by a trainee (i.e., the inter-observer variation). Bottom two rows: we compare the biomarkers computed by an expert evaluating the same image twice (i.e., the intra-observer variation).

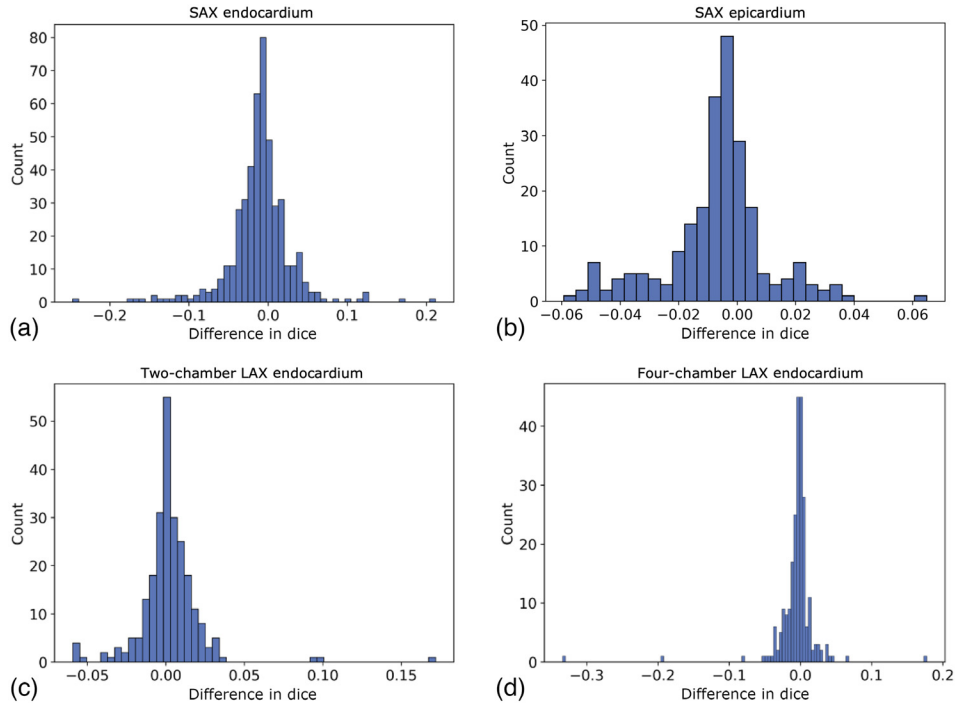

Fig. 9 Histograms showing the difference in Dice coefficients achieved by the fully supervised networks and proposed-100 networks for each segmentation target: (a) SAX endocardium, (b) SAX epicardium, (c) 2CH LAX endocardium, and (d) 4CH LAX endocardium.

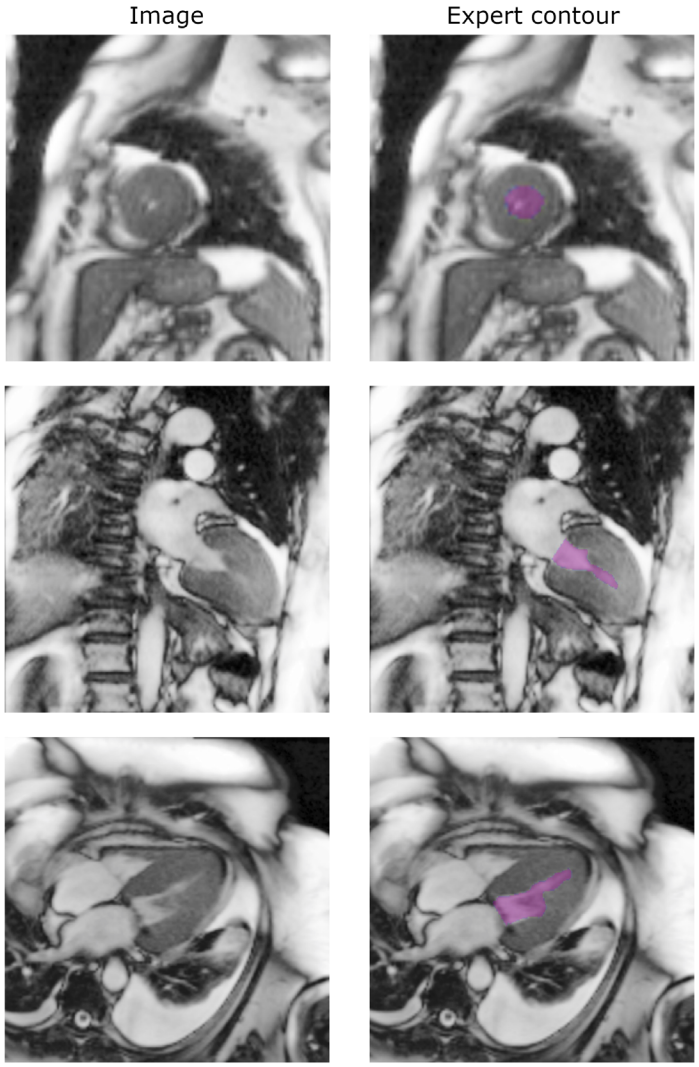

Fig. 10 Example patients with small left ventricular cavities. In panel (a), we show the image only and in panel (b), we show the segmentation mask annotated by an expert.

Table 6 Cardiac functional biomarker values.

	EF	EDV	ESV	SV	LVM
Proposed-100	0.58 (0.12)	168 (54)	74 (48)	94 (26)	132 (43)
Trainee	0.61 (0.13)	157 (52)	64 (46)	92 (26)	146 (44)
Expert	0.59 (0.12)	160 (51)	69 (45)	91 (23)	142 (44)

The mean (standard deviation) values for each cardiac functional biomarker derived from proposed-100 and measured by the trainee and expert annotator on the scan-rescan dataset.

sampling). Here, we show the impact of including the uncertainty-based random sampling scheme when using both supervised and self-supervised losses. We compare DA+CL (each of which use uncertainty-based random sampling for selecting DAs) against naïve DA+CL (neither of which use uncertainty-based random sampling for selecting DAs). We show the results in Table 10.

Table 7 Error mode evaluation.

		SAX endocardium	SAX epicardium	2CH LAX endocardium	4CH LAX endocardium
Proposed-100	Patients with low ESV	0.856 (0.114)	0.941 (0.028)	0.938 (0.032)	0.922 (0.058)
	Patients with high ESV	0.924 (0.034)	0.950 (0.017)	0.967 (0.021)	0.953 (0.034)
Fully supervised	Patients with low ESV	0.864 (0.101)	0.947 (0.026)	0.937 (0.032)	0.925 (0.060)
	Patients with high ESV	0.937 (0.031)	0.951 (0.022)	0.956 (0.044)	0.959 (.029)

The mean (standard deviation) Dice coefficients of proposed-100 and the fully supervised networks on the patients that have the 10% highest and 10% lowest ESV in the test set of the contoured dataset. We report the stratified Dice coefficient for each segmentation target.

Table 8 Ablations.

	10 slices	50 slices	100 slices	250 slices	500 slices	1000 slices	All 16,812 slices
Naive	0.821 (0.099)	0.868 (0.081)	0.877 (0.084)	0.891 (0.082)	0.897 (0.071)	0.901 (0.069)	0.907 (0.069)
DA	0.827 (0.096)	0.874 (0.081)	0.882 (0.078)	0.893 (0.082)	0.897 (0.072)	0.903 (0.066)	0.909 (0.065)
DA + CL	0.851 (0.087)	0.877 (0.08)	0.882 (0.081)	0.895 (0.077)	0.899 (0.07)	0.900 (0.073)	0.909 (0.066)
DA + CL + PL	0.872 (0.077)	0.889 (0.078)	0.897 (0.072)	0.900 (0.078)	0.902 (0.07)	0.900 (0.071)	0.909 (0.066)

Mean (standard deviation) Dice coefficients for our proposed method and each ablation training method for the SAX cine endocardium segmentation task using different amounts of labeled data, as visualized in main text Fig. 5.

Table 9 Sampling strategies.

	Number of unique patients in labeled dataset	Number of 2D image slices in labeled dataset	Dice coefficient on SAX endocardium segmentation task achieved with proposed method
Sampling strategy 1: select one image slice from each patient	100	100	0.897 (0.072)
Sampling strategy 2: include all image slices from each patient	9	98	0.896 (0.075)

Results describing the impact of different sampling strategies in our proposed pipeline. The mean (standard deviation) Dice coefficient is given for each of the sampling strategies for the SAX endocardium segmentation task test dataset.

Table 10 Uncertainty-based random sampling.

	Proposed-10	Proposed-100	Proposed-1000
DA + CL, using uncertainty-based random sampling	0.851 (0.087)	0.882 (0.081)	0.900 (0.073)
Naïve DA + CL, not using uncertainty-based random sampling	0.841 (0.096)	0.884 (0.075)	0.892 (0.077)

Results describing the impact of using the uncertainty-based random sampling approach to selecting augmentations in both the supervised cross entropy loss and self-supervised CL. The mean (standard deviation) Dice coefficient is given for the SAX endocardium segmentation task test dataset using the proposed method and 10 (proposed-10), 100 (proposed-100), and 1000 (proposed-1000) labeled image slices.

6.2.4 *Baseline comparisons*

We further compare our proposed method against three previously published semi- and self-supervised methods, each of which rely on unlabeled and labeled data during training.

- Shape-aware semi-supervised 3D semantic segmentation (SASSNet).[7] This semi-supervised approach trains segmentation networks using a supervised loss over a small amount of labeled data and a self-supervised loss over additional unlabeled data to enforce a geometric shape constraint on the segmentation masks output from the neural network.
- Semi-supervised segmentation through dual-task consistency (DTC).[8] This semi-supervised approach uses two supervised loss functions over a small amount of labeled data and a self-supervised loss over additional unlabeled data to enforce task consistency between different task heads.
- Contrastive learning for medical image segmentation (SimCLRSeg).[10] This is a self-supervised approach that first pretrains a backbone using a contrastive loss designed specifically for medical image segmentation, then fine tunes the backbone using a small amount of labeled data.

For all methods we use a U-Net architecture. We use the codebases provided by the authors of each baseline and perform a hyperparameter grid search for each method and segmentation task. For each approach, we use the same five image volumes as labeled data and the remainder of the training set as unlabeled training data. We also evaluate the fully supervised U-Net performance, which is trained using all labeled data, as an upper bound on performance.

In Table 11, we show the mean Dice coefficients over the test sets for the CMR SAX segmentation tasks. We show the performance of five different training methods: three previously published approaches, the approach proposed in this work, and the fully supervised network.

Table 11 Baselines.

	CMR SAX endocardium	CMR SAX epicardium
SASSNet	0.625 (0.176)	0.693 (0.152)
DTC	0.582 (0.232)	0.679 (0.154)
SimCLRSeg	0.841 (0.108)	0.921 (0.047)
Proposed	0.870 (0.088)	0.928 (0.043)
Fully supervised	0.899 (0.08)	0.949 (0.043)
Labeled training data reduction	99.6%	99.4%

Mean Dice coefficients (standard deviation) over the test set for the CMR SAX endocardium and epicardium tasks, trained with five labeled patients' data. We show the performance of the proposed method, three previously proposed approaches, and the fully supervised network. We also report the training dataset labeling reduction achieved for each task compared to the corresponding fully supervised network.

We also report the training dataset labeling reduction achieved for each task compared to the corresponding fully supervised network. We observe that the proposed method exceeds the performance of the previously proposed approaches.

Disclosures

Erik Schelbert is on the scientific advisory board for Haya Therapeutics. The remaining authors report no conflicts of interest.

Acknowledgments

Sarah Hooper is supported by the Fannie and John Hertz Foundation, the National Science Foundation Graduate Research Fellowship (Grant No. DGE-1656518), and as a Texas Instruments Fellow under the Stanford Graduate Fellowship in Science and Engineering. Curtis Langlotz is supported by the Medical Imaging Data Resource Center (MIDRC), funded by the National Institute of Biomedical Imaging and Bioengineering (NIBIB) of the National Institutes of Health (NIH, Grant Nos. 75N92020C00008 and 75N92020C00021). This material is based upon work supported by the Google Cloud Research Credits program with the award GCP19980904. We gratefully acknowledge the support of NIH (Grant No. U54EB020405) (Mobilize) and NSF [Grant Nos. CCF1763315 (Beyond Sparsity), CCF1563078 (Volume to Velocity), and 1937301 (RTML)]; ONR (Grant No. N000141712266) (Unifying Weak Supervision); ONR (Grant No. N00014-20-1-2480): Understanding and Applying Non-Euclidean Geometry in Machine Learning; [Grant No. N000142012275 (NEPTUNE)]; and the Moore Foundation, NXP, Xilinx, LETI-CEA, Intel, IBM, Microsoft, NEC, Toshiba, TSMC, ARM, Hitachi, BASF, Accenture, Ericsson, Qualcomm, Analog Devices, the Okawa Foundation, American Family Insurance, Google Cloud, Salesforce, Total, the HAI-Google Cloud Platform Cloud Credits for Research program, the Stanford Data Science Initiative (SDSI), and members of the Stanford DAWN project: Facebook, Google, and VMWare. The Mobilize Center is a Biomedical Technology Resource Center, funded by the NIH NIBIB (Grant No. P41EB027060). The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views, policies, or endorsements, either expressed or implied, of NIH, ONR, or the U.S. Government.

References

1. G. Litjens et al., "A survey on deep learning in medical image analysis," *Med. Image Anal.* **42**, 60–88 (2017).
2. M. H. Hesamian et al., "Deep learning techniques for medical image segmentation: achievements and challenges," *J. Digit. Imaging* **32**(4), 582–596 (2019).
3. D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," *Annua. Rev. Biomed. Eng.* **19**, 221–248 (2017).
4. A. N. Bhuva et al., "A multicenter, scan-rescan, human and machine learning CMR study to test generalizability and precision in imaging biomarker analysis," *Circulation Cardiovasc. Imaging* **12**(10), e009214 (2019).
5. W. Bai et al., "Semi-supervised learning for network-based cardiac MR image segmentation," *Lect. Notes Comput. Sci.* **10434**, 253–260.
6. G. Bortsova et al., "Semi-supervised medical image segmentation via learning consistency under transformations," *Lect. Notes Comput. Sci.* **11769**, 810–818.
7. S. Li, C. Zhang, and X. He, "Shape-aware semi-supervised 3d semantic segmentation for medical images," *Lect. Notes Comput. Sci.* **12261**, 552–561 (2020).
8. X. Luo et al., "Semi-supervised medical image segmentation through dual-task consistency," in *Proc. AAAI Conf. Artif. Intell.*, Vol. 35, pp. 8801–8809 (2021).

9. C. Ouyang et al., "Self-supervision with superpixels: training few-shot medical image segmentation without annotation," *Lect. Notes Comput. Sci.* **12374**, 762–780.
10. K. Chaitanya et al., "Contrastive learning of global and local features for medical image segmentation with limited annotations," in *Proc. 34th Int. Conf. Neural Inf. Process. Syst.*, pp. 12546–12558 (2020).
11. K. Alfakih et al., "Assessment of ventricular function and mass by cardiac magnetic resonance imaging," *Eur. Radiol.* **14**(10), 1813–1822 (2004).
12. H. D. White et al., "Left ventricular end-systolic volume as the major determinant of survival after recovery from myocardial infarction," *Circulation* **76**(1), 44–51 (1987).
13. C. Chen et al., "Deep learning for cardiac image segmentation: a review," *Front. Cardiovasc. Med.* **7**, 25 (2020).
14. F. Isensee et al., "Automatic cardiac disease assessment on cine-MRI via time-series segmentation and domain specific features," *Lect. Notes Comput. Sci.* **10663**, 120–129.
15. N. Painchaud et al., "Cardiac segmentation with strong anatomical guarantees," *IEEE Trans. Med. Imaging* **39**(11), 3703–3713 (2020).
16. O. Bernard et al., "Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved?" *IEEE Trans. Med. Imaging* **37**(11), 2514–2525 (2018).
17. T. Leiner et al., "Machine learning in cardiovascular magnetic resonance: basic concepts and applications," *J. Cardiovasc. Magn. Reson.* **21**(1), 1–14 (2019).
18. G. Litjens et al., "State-of-the-art deep learning in cardiovascular image analysis," *JACC: Cardiovasc. Imaging* **12**(8 Part 1), 1549–1565 (2019).
19. D. M. Vigneault et al., "Ω-net (omega-net): fully automatic, multi-view cardiac MR detection, orientation, and segmentation with deep neural networks," *Med. Image Anal.* **48**, 95–106 (2018).
20. S. Leng et al., "Computational platform based on deep learning for segmenting ventricular endocardium in long-axis cardiac MR imaging," in *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, pp. 4500–4503 (2018).
21. W. Bai et al., "Automated cardiovascular magnetic resonance image analysis with fully convolutional networks," *J. Cardiovasc. Magn. Reson.* **20**(1), 65 (2018).
22. H. Xue et al., "Automated inline analysis of myocardial perfusion MRI with deep learning," *Radiol. Artif. Intell.* **2**(6), e200009 (2020).
23. E. Puyol-Antón et al., "Automated quantification of myocardial tissue characteristics from native T1 mapping using neural networks with uncertainty-based quality-control," *J. Cardiovasc. Magn. Reson.* **22**(1), 1–15 (2020).
24. S. Moccia et al., "Development and testing of a deep learning-based strategy for scar segmentation on CMR-LGE images," *Magn. Reson. Mater. Phys. Biol. Med.* **32**(2), 187–195 (2019).
25. O. Ronneberger, P. Fischer, and T. Brox, "U-net: convolutional networks for biomedical image segmentation," *Lect. Notes Comput. Sci.* **9351**, 234–241.
26. Ö. Çiçek et al., "3D U-Net: learning dense volumetric segmentation from sparse annotation." *Lect. Notes Comput. Sci.* **9901**, 424–432.
27. A. J. Ratner et al., "Learning to compose domain-specific transformations for data augmentation," in *Adv. Neural Inf. Process. Syst.*, Vol. **30**, p. 3239 (2017).
28. E. D. Cubuk et al., "AutoAugment: learning augmentation strategies from data," in *IEEE/CVF Conf. Comput. Vis. and Pattern Recognit. (CVPR)*, pp. 113–123 (2019).
29. S. Wu et al., "On the generalization effects of linear transformations in data augmentation," in *Proc. 37th Int. Conf. Mach. Learn.*, pp. 10410–10420 (2020).
30. A. J. Ratner et al., "Data programming: creating large training sets, quickly," in *Adv. Neural Inf. Process. Syst.*, Vol. 29, pp. 3567–3575 (2016).
31. A. Ratner et al., "Snorkel: rapid training data creation with weak supervision," in *Proc. VLDB Endowment*, Vol. 11, p. 269.
32. D.-H. Lee, "Pseudo-label: the simple and efficient semi-supervised learning method for deep neural networks," in *Workshop on Challenges in Represent. Learn., ICML*, Vol. **3**, p. 896 (2013).

33. B. Zoph et al., "Rethinking pre-training and self-training," arXiv:2006.06882 (2020).
34. T. Chen et al., "A simple framework for contrastive learning of visual representations," in *Proc. 37th Int. Conf. Mach. Learn.*, pp. 1597–1607 (2020).
35. K. He et al., "Momentum contrast for unsupervised visual representation learning," pp. 9729–9738 (2020).
36. Y. Tian et al., "What makes for good views for contrastive learning?" arXiv:2005.10243 (2020).
37. M. Caron, et al., "Unsupervised learning of visual features by contrasting cluster assignments," arXiv:2006.09882 (2020).
38. X. Zhao et al., "Contrastive learning for label efficient semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vision*, pp. 10623–10633 (2021).
39. J.-B. Grill et al., "Bootstrap your own latent: a new approach to self-supervised learning," arXiv:2006.07733 (2020).
40. M. Caron et al., "Emerging properties in self-supervised vision transformers," arXiv:2104.14294 (2021).
41. D. Berthelot et al., "Mixmatch: a holistic approach to semi-supervised learning," arXiv:1905.02249 (2019).
42. D. Berthelot et al., "Remixmatch: semi-supervised learning with distribution alignment and augmentation anchoring," arXiv:1911.09785 (2019).
43. K. Sohn et al., "Fixmatch: simplifying semi-supervised learning with consistency and confidence," arXiv:2001.07685 (2020).
44. A. Tarvainen and H. Valpola, "Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results," arXiv:1703.01780 (2017).
45. Q. Xie et al., "Unsupervised data augmentation for consistency training," arXiv:1904.12848 (2019).
46. C.-W. Kuo et al., "Featmatch: feature-based augmentation for semi-supervised learning," *Lect. Notes Comput. Sci.* **12363**, 479–495.
47. Y. Zou et al., "Pseudoseg: designing pseudo labels for semantic segmentation," arXiv:2010.09713 (2020).
48. M. Assran et al., "Semi-supervised learning of visual features by non-parametrically predicting view assignments with support samples," arXiv:2104.13963 (2021).
49. S. Hooper et al., "Cut out the annotator, keep the cutout: better segmentation with weak supervision," OpenReview.net (2020).
50. H. Roth et al., "Cardiac segmentation of LGE MRI with noisy labels," *Lect. Notes Comput. Sci.* **12009**, 228–236.
51. S. Wang et al., "Annotation-efficient deep learning for automatic medical image segmentation," *Nat. Commun.* **12**, 5915 (2021).
52. R. H. Davies et al., "Precision measurement of cardiac structure and function in cardiac MR using machine learning," *J. Cardiovasc. Magn. Reson.* **24**(1), 16 (2022).
53. T. A. Treibel et al., "Extracellular volume associates with outcomes more strongly than native or post-contrast myocardial T1," *Cardiovasc. Imaging* **13**(1_Part_1), 44–54 (2020).
54. A. Bhuva, "The VOLUMES resource," 2019, https://thevolumesresource.com/.
55. M. S. Hansen and T. S. Sørensen, "Gadgetron: an open source framework for medical image reconstruction," *Magn. Reson. Med.* **69**(6), 1768–1776 (2013).
56. H. Xue et al., "Distributed MRI reconstruction using gadgetron-based cloud computing," *Magn. Reson. Med.* **73**(3), 1015–1025 (2015).
57. H. Xue, "CMR landmark detection," 2020, https://github.com/xueh2/CMR_LandMark_Detection/blob/master/models/resunet.py.
58. M. L. Waskom, "Seaborn: statistical data visualization," *J. Open Source Softw.* **6**(60), 3021 (2021).
59. I. Lawrence and K. Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics* **45**(1), 255–268 (1989).
60. N. P. Hyslop and W. H. White, "Estimating precision using duplicate measurements," *J. Air Waste Manage. Assoc.* **59**(9), 1032–1039 (2009).

61. S. Seabold and J. Perktold, "Statsmodels: econometric and statistical modeling with Python," in *Proc. 9th Python in Sci. Conf.*, p. 57 (2010).
62. A. Subbaswamy and S. Saria, "From development to deployment: dataset shift, causality, and shift-stable models in health AI," *Biostatistics* **21**(2), 345–352 (2020).
63. V. Antun et al., "On instabilities of deep learning in image reconstruction and the potential costs of AI," *Proc. Natl. Acad. Sci.* **117**(48), 30088–30095 (2020).
64. J. R. Zech et al., "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study," *PLoS Med.* **15**(11), e1002683 (2018).
65. X. Wang et al., "Inconsistent performance of deep learning models on mammogram classification," *J. Am. Coll. Radiol.* **17**(6), 796–803 (2020).
66. Y. Chen et al., "Towards to robust and generalized medical image segmentation framework."
67. J. De Fauw et al., "Clinically applicable deep learning for diagnosis and referral in retinal disease," *Nat. Med.* **24**(9), 1342–1350 (2018).
68. B. C. Lowekamp et al., "The design of SimpleITK," *Front. Neuroinf.* **7**, 45 (2013).

Sarah M. Hooper received her BS degree in electrical engineering and a minor in global health technologies from Rice University in 2017 and her MS degree in electrical engineering from Stanford University in 2020. Currently, she is an electrical engineering PhD candidate at Stanford University. Her research interests lie at the intersection of machine learning and medical imaging with a focus on developing new technical tools to increase the accessibility and quality of healthcare.

Biographies of the other authors are not available.