Integrating Deep Metric Learning with Coreset for Active Learning in 3D Segmentation

Arvind Murari Vepa
UCLA
amvepa@ucla.edu

Zukang Yang
UC Berkeley
zukangy@berkeley.edu

Andrew Choi
Horizon Robotics
asjchoi@ucla.edu

Jungseock Joo
UCLA
jjoo@comm.ucla.edu

Fabien Scalzo
UCLA
fab@cs.ucla.edu

Yizhou Sun
UCLA
yzsun@cs.ucla.edu

Abstract

Deep learning has seen remarkable advancements in machine learning, yet it often demands extensive annotated data. Tasks like 3D semantic segmentation impose a substantial annotation burden, especially in domains like medicine, where expert annotations drive up the cost. Active learning (AL) holds great potential to alleviate this annotation burden in 3D medical segmentation. The majority of existing AL methods, however, are not tailored to the medical domain. While weakly-supervised methods have been explored to reduce annotation burden, the fusion of AL with weak supervision remains unexplored, despite its potential to significantly reduce annotation costs. Additionally, there is little focus on slice-based AL for 3D segmentation, which can also significantly reduce costs in comparison to conventional volume-based AL. This paper introduces a novel metric learning method for Coreset to perform slice-based active learning in 3D medical segmentation. By merging contrastive learning with inherent data groupings in medical imaging, we learn a metric that emphasizes the relevant differences in samples for training 3D medical segmentation models. We perform comprehensive evaluations using both weak and full annotations across four datasets (medical and non-medical). Our findings demonstrate that our approach surpasses existing active learning techniques on both weak and full annotations and obtains superior performance with low-annotation budgets which is crucial in medical imaging. Source code for this project is available in the supplementary materials and on GitHub: `https://github.com/arvindmvepa/al-seg`.

1 Introduction

In the field of 3D medical segmentation, manual annotation of entire volumes, despite being laborious and time-consuming, has been the gold standard. Annotating a single 2D image can take minutes to hours depending on the complexity of the image (75; 73; 15; 71; 6), and a 3D medical volume, containing up to 200 slices, can require a significant amount of expert labor. Annotating a full dataset not only imposes a significant time burden on medical experts but also incurs high costs. Therefore, active learning (AL) is urgently needed to optimize annotation efforts.

Surprisingly, the potential of AL in the context of 3D medical segmentation has not been extensively explored. Traditional AL techniques typically focus on either diversity or model uncertainty, often neglecting relevant groupings within the data. For example, slices from the same patient or volume tend to show consistent characteristics. We propose a deep metric learning strategy that identifies and utilizes these similarities to better highlight diversity in the active learning process. Diverse samples

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

help the model learn a wide variety of patterns and features, which can be crucial for generalization. While medical imaging has natural groupings which we can leverage, our approach extends to a wide range of real-world datasets, including video segmentation.

Several other strategies may also help in reducing costs. Most current AL approaches use volume-based AL (39) rather than slice-based AL for 3D segmentation, which tends to be more costly and less efficient. Alternatively, weakly supervised methods, which require simpler annotations like scribbles (34; 29; 11), bounding boxes (13; 83), points (82), or semi-automated techniques (5; 44; 62; 56), have been shown to perform comparably to fully supervised methods (72; 40; 54; 13; 26; 59). However, combining AL with weak supervision, especially in medical settings, remains unexplored. In our research, we explore both slice-based AL and the integration of AL with weak supervision as potential cost-reducing measures.

In this paper, we present our contributions to AL for 3D medical segmentation:

1. The first work to integrate deep metric learning with Coreset during active learning for 3D medical segmentation. Our approach shows superior performance across four datasets (medical and non-medical) with low annotation budgets.
2. The first work to comprehensively compare new and existing algorithms for slice-based active learning for 3D medical segmentation utilizing both weak and full annotations.

2 Related work

Active learning AL methods can be broadly classified into 1) uncertainty-based and 2) diversity-based methods (76). Uncertainty-based methods include deep bayesian methods (21; 45; 38; 65), deep ensembles (3; 14; 52; 32; 20), contrastive learning methods (85; 41; 87; 36), and geometry-based methods (19). Diversity-based methods include coreset-based methods (64; 7), clustering-based methods (23), variational adversarial learning methods (66; 37), and random sampling methods (48). A limitation of previous methods is that they are too general and fail to utilize common data groupings found in real-world datasets. Prior methods (67; 9; 69) have tried to incorporate groupings but do not utilize domain information to generate them, which can be suboptimal. Recent research has started integrating domain-specific data groupings into AL algorithms with some success(28; 79), but these methods are designed for specific uses and lack broad applicability. While other methods have adapted Coreset (31; 30), they are not specifically tailored to 3D medical segmentation as our method is.

Deep metric learning Metric learning is focused on developing methods to measure similarity between data points and used in many applications. Recently, metric learning has focused on deep learning-based feature representations for data points (46). Contrastive losses are popular for metric learning, including Triplet Loss (63) and NT-Xent loss (67). Several non-contrastive approaches have been proposed based on center loss (77; 17; 16), proxy-based methods (47; 70; 27), and LLM guidance (61). One interesting approach is ensemble deep metric learning which combines embeddings from an ensemble of encoders (1; 43; 53; 60; 80). Recent work has improved on ensemble-based methods by factorizing the network training based on differerent objectives (74). However, these approaches can be computationally expensive and narrowly tailored to specific applications. Additionally, non-contrastive approaches often require class supervision to perform well. While some non-contrastive approaches do not (88; 18), they are also narrowly tailored to specific applications.

In contrast, contrastive learning is often used in self-supervised settings in diverse applications (10; 2; 55). The NT-Xent loss specifically outperformed other contrastive losses in self-supervised zero-shot image classification and outperformed fully-supervised ResNet-50 (12). However, prior work (including in active learning (85; 23; 39)) does not leverage any inherent data groupings in the contrastive loss, which can be useful weak supervision. Additionally, recent work in active learning can be computationally expensive because they retrain the feature encoder after each active learning round (85).

Active learning in medical segmentation There has been some notable research on AL in medical segmentation. Earlier work utilized bootstrapping to estimate sample uncertainty (81). In another work, the researchers built a mutual information-based metric between the labeled and unlabeled pools to improve diversity (49). However, both of these approaches are computationally expensive

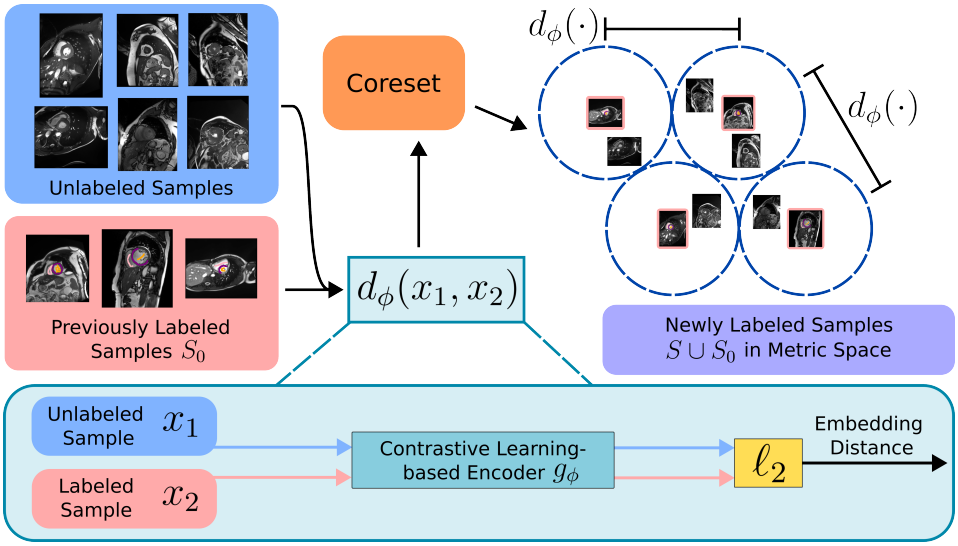


Figure 1: Overview of our active learning pipeline

and not scalable for large datasets. There were also inefficiencies in previous studies, such as choosing whole volumes instead of individual slices in 3D segmentation, which increased costs (39). Random sampling proved effective in some cases (6) and calculating uncertainty using stochastic batches was also effective (20). None of the prior approaches leverage the data groupings inherent in medical 3D data and also do not focus on active learning with weak annotations (e.g., scribbles). While another method also utilizes groupings in medical data (86), their groupings are assumed to be quite large and it's unclear how they would extend their group classification approach to when there are a large number of patients, volumes, and adjacent slice groups like with our datasets.

3 Methodology

3.1 Problem Formulation

In this section, we formally describe the problem of active learning for 3D segmentation. Let $\mathcal{X} \subset \mathbb{R}^{h \times w \times d}$ be the set of 3D volumes and $\mathcal{Y} \subset \{0,1\}^{h \times w \times d \times k}$ be the set of 3D masks where h, w, d correspond to the height, width, and depth (number of slices) of a 3D volume and k refers to the number of classes. In 3D segmentation, we learn a mapping $F : \mathcal{X} \to \mathcal{Y}$.

Consider a loss function $\mathcal{L} : \mathcal{F} \times \mathcal{Y} \to \mathbb{R}$ where $\mathcal{F}$ is the range of model prediction probabilities ($\mathcal{F} = [0,1]^{h \times w \times d \times k}$). We consider the dataset D a large collection of data points which are sampled $i.i.d.$ over the space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ as $\{\mathbf{x}_i, \mathbf{y}_i\}_{i \in [n]} \sim p_{\mathcal{Z}}$. We additionally consider a partially labeled subset $s \subset D$. Thus, active learning for 3D segmentation can be formulated as follows:

$$\underset{\Delta s \subset D, |\Delta s| \leq k}{\operatorname{argmin}} \; E_{\mathbf{x},\mathbf{y} \sim p_{\mathcal{Z}}}[\mathcal{L}(\hat{\mathbf{y}}, \mathbf{y})] \tag{1}$$

where $\hat{\mathbf{y}}$ are the model predictions $F(V)$ (where $V = [\text{slice}_1, ..., \text{slice}_d]$ is the volume), Δs is the optimal requested labeled set, k is the active learning budget, and F is the deep learning method learned from $s \cup \Delta s$. The Coreset approach is one method of solving Equation 1 (64). However, there are two distinguishing factors in our current work: 1) slice-based active learning for 3D segmentation and 2) deep metric learning. In slice-based 3D segmentation, we learn a mapping f from $\mathcal{X}' \to \mathcal{Y}'$ where $\mathcal{X}' \subset \mathbb{R}^{h \times w}$, $\mathcal{Y}' \subset \{0,1\}^{h \times w \times k}$, and $F(V) = [f(\text{slice}_1), ..., f(\text{slice}_d)]$. Additionally, D is a collection of slices which are sampled $i.i.d.$ over the space $\mathcal{Z}' = \mathcal{X}' \times \mathcal{Y}'$ as $\{\mathbf{x}'_i, \mathbf{y}'_i\}_{i \in [n]} \sim p_{\mathcal{Z}'}$.

In the original Coreset paper (64), $s \cup \Delta s$ is the set cover of D with radius δ. However, the Euclidean metric used to calculate the radius does not utilize task-relevant information and may be sub-optimal. This motivates us to consider deep metric learning to utilize more task-relevant information, where $d_\phi(x_1, x_2)$ is some parameterized metric. In the the original paper (64), the authors provide a theorem

which bounds the loss from Equation 1 based on the radius δ. We show a very similar bound where $\delta = \max_{x_1 \in D} \min_{x_2 \in s \cup \Delta s} d_\phi(x_1, x_2)$ (we defer the precise bound and proof to the Appendix B). This leads to our Coreset optimization formulation:

$$\operatorname*{argmin}_{\Delta s \subset D, |\Delta s| \leq k} \max_{x_1 \in D} \min_{x_2 \in s \cup \Delta s} d_\phi(x_1, x_2) \tag{2}$$

The above formulation can be intuitively defined as follows; choose k additional center points such that the largest distance between a data point and its nearest center is minimized (64).

3.2 Metric learning

The parameterized metric $d_\phi(x_1, x_2)$ is defined as:

$$d_\phi(x_1, x_2) = \ell_2(g_\phi(x_1), g_\phi(x_2)) \tag{3}$$

where ℓ_2 is the Euclidean metric. Our goal is to learn g_ϕ, or a feature representation, that emphasizes task-relevant similarities and differences in the data for selecting diverse samples for Coreset. We do this by training a contrastive learning-based encoder with a unique Group-based Contrastive Learning (GCL) which utilizes inherent data groupings specific to medical imaging to generate the feature representation. d_ϕ is then used by the Coreset algorithm to select the optimal set of slices. A flow chart illustrating our pipeline can be seen in Figure 1. The annotations for the collected slices are then used to train a segmentation model.

3.2.1 Group-based Contrastive Learning for feature representation in metric learning

While there may be several task-relevant groupings in 3D medical imaging, it is not immediately apparent which of these would be useful for feature representation for metric learning for Coreset. For the ACDC dataset, we note the mean pairwise absolute deviation of the normalized training slice pixel values within different groups averaged over the dataset: 0.217 over the entire dataset, 0.159 within patient groups, 0.166 within volume groups, and 0.115 within adjacent slice groups. Note that the volume groups and the adjacent slice groups are the most and least diverse respectively. While intuitively the most diverse group would have the most important features for diversity, this does not necessarily indicate what combinations of groups would be helpful as well.

Instead, we propose a general group contrastive loss based on NT-Xent loss (67). The NT-Xent loss focuses on generating and comparing embeddings for image pairs and their augmentations. It promotes similarity in embeddings for the same image and its augmentation while encouraging dissimilarity for different images. In the same vein, in our group-based loss, we promote similar embeddings for slices from the same group and dissimilar embeddings for slices from different groups, enhancing group-level representation. We define "group" as a set of 2D slices associated with one patient. The formula is as follows:

$$\mathcal{L}_{\text{group}} = -\frac{1}{NG} \sum_{i=1}^{N} \sum_{j \neq i, g_j = g_i}^{N} \log \frac{\exp(sim(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[(k \neq i) \wedge ((g_k = g_i) \vee (p_k \neq p_i))]} \exp(sim(z_i, z_k)/\tau)} \tag{4}$$

in which $i \in \{1, 2, ..., N\}$ are the indices for the standard batch slices, $j \in \{N+1, N+2, ..., 2N\}$ are the indices for the augmented batch slices, g_i refers to the group associated with slice i in the batch, p_i refers to the patient associated with slice i in the batch, G represents the average number of slices per group (calculated as the total number of unaugmented slices divided by the total number of groups), z is the embedding for a slice in the batch, τ is a temperature parameter, and sim is a similarity function which was cosine similarity in this study. The formula shares similarities with NT-Xent loss but introduces some modifications:

1. There are two summations which reflects all the group slices for a particular data point.
2. In the numerator of the logarithmic term, all the group slices for particular data point are considered similar, encouraging the development of similar embeddings for these slices.

3. The denominator excludes patient slices for a particular data point that are not part of the same group

Excluding non-group patient slices ensures that the model does not promote dissimilarity between non-group slices from the same patient. This ensures that we can sum multiple group losses together without counteracting their effects. In our study, we considered patient, volume, adjacent slice group contrastive losses in addition to NT-Xent loss. Our overall loss formulation is as follows:

$$\mathcal{L}_{\text{contrastive}} = \mathcal{L}_{\text{NT-Xent}} + \lambda_1 \mathcal{L}_{\text{patient group}} + \lambda_2 \mathcal{L}_{\text{volume group}} + \lambda_3 \mathcal{L}_{\text{slice group}} \tag{5}$$

where $\mathcal{L}_{\text{patient group}}, \mathcal{L}_{\text{volume group}}, \mathcal{L}_{\text{slice group}}$ are the group contrastive losses associated with patient, volume, and adjacent slice groups respectively and $\{\lambda_1, \lambda_2, \lambda_3\}$ are constants. The summation of different contrastive losses in our overall loss $\mathcal{L}_{\text{contrastive}}$ is the focus of the ablation study in Section 4.6.

Batch sampler We employ $\mathcal{L}_{\text{contrastive}}$ to train a SimCLR network (35) using a ResNet-18 encoder (24), which is our g_ϕ. In practice, standard random data loading would yield minimal impact from $\mathcal{L}_{\text{contrastive}}$ due to the low probability of randomly selecting two slices from the same group (e.g., same patient, same volume, or adjacent slices) within a batch. To address this, we introduce a batch sampler designed to increase their occurrence. The batch sampling process can be summarized as follows: 1. create a singular slice group for all the dataset slices, 2. for each group, randomly include a slice from the same patient for each of the groupings used (e.g., patient and volume), 3. combine groups from different patients to form a batch. An epoch consists of all the dataset groups: thus, more contrastive loss groups result in larger epochs. The batch sampler would be reset every epoch to ensure randomness during training. Please see Appendix C for more details.

Training the SimCLR network with this batch sampler eliminates the need for complex algorithmic adjustments to accommodate multiple contrastive loss groups, a challenge in other AL contrastive learning methodologies (85). We conduct the training over 100 epochs using an ADAM optimizer, with a learning rate of 3e-4, a weight decay of 1.0e-6, and a batch size of 8 for one or three groups and 9 for two groups.

3.2.2 Segmentation model training

Once we have g_ϕ, we can calculate d_ϕ and collect annotations for unlabeled slices to train our segmentation model. Our AL evaluation consists of several rounds of annotation collection from an oracle. After each round of annotation collection, we train five segmentation models and record the model's test score with the highest validation score. We repeat the AL experiment five times for each algorithm, each time with a different random seed, and report the average model test score per round. For weakly-supervised segmentation, we train a Dual-Branch Network with Dynamically Mixed Pseudo Labels Supervision (DMPLS) (40), which reported strong metrics on the ACDC dataset using weak supervision. For full supervision, we train UNet (58) which is a frequently used fully-supervised segmentation baseline model. We calculate the 3D DICE score on each 3D volume and report the average on all the volumes in the test set. For full supervision, we also provide results using a pre-trained segmentation model with a ResNet-50 backbone (42) (pre-trained on medical images (42) for medical datasets and ImageNetV2 for non-medical datasets). We do not report weakly-supervised results on pre-trained architectures because the weakly-supervised architectures cannot easily utilize pre-trained backbones.

4 Experiments

4.1 Datasets

ACDC (Automated Cardiac Diagnosis Challenge) The ACDC dataset (4) consists of 200 short-axis cine-MRI scans from 100 patients in the training set and 100 scans from 50 patients in the test set. Acquired data was fully anonymized and handled within the regulations set by the local ethical committee of the Hospital of Dijon (France). Each patient has two annotated end-diastolic (ED) and end-systolic (ES) phase scans. Annotations are available for three structures: the right ventricle (RV), myocardium (Myo), and left ventricle (LV). Additionally, scribble annotations have been provided for each scan by a previous study (72). The training set size consists of 1448 slices.

CHAOS (Combined Healthy Abdominal Organ Segmentation) The CHAOS dataset (33) comprises of abdominal CT images from 20 subjects in the training set, primarily used for liver and vessel segmentation. The anonymized dataset was collected from the Department of Radiology, Dokuz Eylul University Hospital, Izmir, Turkey and the study was approved by the Institutional Review Board of Dokuz Eylul University. Each patient's CT scans contains approximately 144 slices. Binary segmentation masks for the liver are provided. We resampled, cropped, and normalized the images, following the process described in a previous study (72). We partitioned the training set into training (75%), validation (10%), and test (15%) subsets. The training set size consists of 2351 slices.

MS-CMR (Multi-sequence Cardiac MR Segmentation Challenge) The MS-CMR dataset (22; 89; 78) contains late gadolinium enhancement (LEG) MRI images from 45 patients who underwent cardiomyopathy. The data has been collected from Shanghai Renji hospital with institutional ethics approval and has been anonymized. These images were multi-slice, acquired in the ventricular short-axis views. We obtained realistic and manual scribble annotations from a prior study (84). Similar to this study (84), we split the data such that 25 patients were assigned to train, 5 to validation, and 20 to test, resulting in 382 slices in the training set. For data processing, we resampled the images into a resolution of 1.37x1.37 mm, and then they were cropped or padded to a fixed size of 212 x 212 pixels. During training, the image pixel values were normalized to zero mean and unit variance.

DAVIS (Densely Annotated Video Segmentation) The DAVIS dataset (50; 51) is a densely annotated video dataset associated with the 2016 DAVIS and 2017 DAVIS Challenge. The dataset collection was partially funded by SNF and human subject data was collected ethically, to the best of our knowledge. In our study, we utilized the train and val sets associated with the 2016 DAVIS Challenge which contained 30 and 20 videos respectively and all frames with 480p resolution. While we used the 2016 DAVIS train set, we created the val and test set by further splitting the 2016 DAVIS val set into 5 and 15 videos respectively. Our train set consisted of 2079 densely annotated frames.

For additional generalizability of our approach, we evaluated our methodology on both weak and full supervision, depending on data availability. Because the CHAOS, MS-CMR, and DAVIS dataset do not contain any hierarchical organization of multiple volumes/videos, we did not use the patient group loss. For video segmentation, in our contrastive loss we treated each video as a "volume" and each video frame as a "slice"; thus, we considered both the volume and adjacent slice group loss.

4.2 Experimental settings

When conducting experiments with pre-trained segmentation models, for ACDC, CHAOS, and MS-CMR we collect annotations in cumulative increments of 2%, 3%, 4%, 5%, 10%, 15%, 20%, and 40% in each round. Because segmentation networks require more training data for natural images, for the DAVIS dataset we collect annotations in cumulative increments of 10%, 20% , 30%, and 40% in each round. When conducting experiments with pre-trained segmentation models, because of the benefit of prior pre-training, we collect annotations in cumulative increments of 1%, 2%, 3%, 4%, and 5% in each round.

Because solving the Coreset problem is NP-Hard, we utilized the K-Center Greedy algorithm for our Coreset implementation (48), which is a $2 - OPT$ solution (64) and produces very competitive results in comparison to other more computationally-intensive solutions. We compared our approach to vanilla Coreset (K-Center Greedy) (64), Random Sampling, CoreGCN (7), TypiClust (23; 39), Stochastic Batches (using Deep Ensembles with Entropy) (20), VAAL (66), Deep Ensembles (utilizing Variance Ratio scoring) (3), and Bayesian Deep Learning (utilizing the BALD score) (21). To ensure a fair comparison, all approaches were evaluated using the same experimental settings.

All of our experiments were primarily conducted with a single Tesla V100 GPU on an internal cluster. Our contrastive learning-based encoder and segmentation models consumed approximately 3 GB of GPU memory while training. The contrastive encoder's training speed was approximately 40 slices/second which resulted in a training time of 100 minutes, 200 minutes, and 300 minutes on ACDC for one, two, and three group losses respectively. One single AL experiment for our method on ACDC (eight rounds with five models trained per round) took approximately 24 hours and used about 400 MB of storage.

Table 1: DICE scores for ACDC, MS-CMR, and CHAOS

	Weakly-supervised					Fully-supervised				
	2%	3%	4%	5%	40%	2%	3%	4%	5%	40%
	ACDC									
BALD (21)	44.8	54.8	61.4	66.2	86.4	53.8	66.0	67.9	71.7	89.3
Variance Ratio (3)	43.3	45.0	54.2	61.4	85.9	52.6	62.2	66.6	69.1	86.9
Random (48)	44.9	45.7	59.2	66.9	86.8	66.3	76.9	79.3	80.1	**90.2**
VAAL (66)	41.8	47.8	66.1	72.1	86.4	63.2	75.2	79.5	81.3	89.9
Coreset (64)	45.2	49.8	68.7	70.1	**86.9**	58.9	69.3	75.7	81.9	**90.2**
TypiClust (23; 39)	44.8	45.6	67.6	73.1	85.7	66.6	75.8	79.5	81.7	89.5
Stochastic Batches (20)	36.9	39.5	53.7	57.4	85.8	60.3	69.8	74.1	75.4	89.3
CoreGCN (7)	40.8	49.3	69.1	74.1	86.3	67.7	74.9	79.0	80.7	89.6
Ours	**52.3**	**59.8**	**73.3**	**76.1**	86.4	**70.9**	**77.4**	**81.6**	**82.5**	**90.2**
	Weakly-supervised					Fully-supervised				
	2%	3%	4%	5%	40%	2%	3%	4%	5%	40%
	MS-CMR					CHAOS				
BALD	37.0	48.9	57.4	60.0	85.8	79.6	80.7	80.4	81.1	95.8
Variance Ratio	41.3	46.8	52.8	55.3	84.3	74.9	72.9	76.6	76.2	92.8
Random	39.7	55.0	61.0	61.5	85.7	**80.7**	81.7	84.2	85.1	96.2
Coreset	28.7	53.6	56.9	58.7	**86.7**	80.0	80.4	81.2	88.1	**96.5**
Stochastic Batches	38.5	**56.2**	59.8	60.5	86.2	77.2	**82.9**	83.8	84.7	96.1
CoreGCN	27.0	48.7	57.1	57.2	85.5	67.8	77.7	77.8	74.4	94.3
Ours	**44.3**	53.3	**63.4**	**63.5**	86.3	80.5	82.5	**85.9**	**90.3**	96.3

Table 2: DICE scores for DAVIS

	Fully-supervised				
	10%	20%	30%	40%	Mean
BALD	**43.6**	42.0	43.4	43.8	43.1
Variance Ratio	36.1	31.2	34.9	40.7	35.6
Random	39.6	40.5	**47.4**	**48.5**	45.5
Coreset	31.7	39.4	42.2	42.1	41.2
Stochastic Batches	40.3	41.5	45.1	47.6	44.7
Ours	42.8	**45.2**	45.5	46.6	**45.8**

4.3 Results

Tables 1 and 2 summarize the results from our experiments on weak and full annotations for the ACDC, MS-CMR, CHAOS, and DAVIS datasets when trained from scratch. Our method excels in low annotation budget scenarios (2%-5%) on the ACDC dataset, outperforming other methods by up to 10% in some cases. This advantage is vital in the medical field where annotation costs are often high. On the MS-CMR, CHAOS, and DAVIS datasets, our method remains highly competitive,

Table 3: Mean DICE scores over all annotation datapoints with pre-trained weights

	Fully-supervised		
	ACDC	CHAOS	DAVIS
Random	78.4	94.9	74.9
Stochastic Batches	77.3	95.0	**75.1**
Coreset	78.6	95.1	74.6
Ours	**79.5**	**95.2**	**75.1**

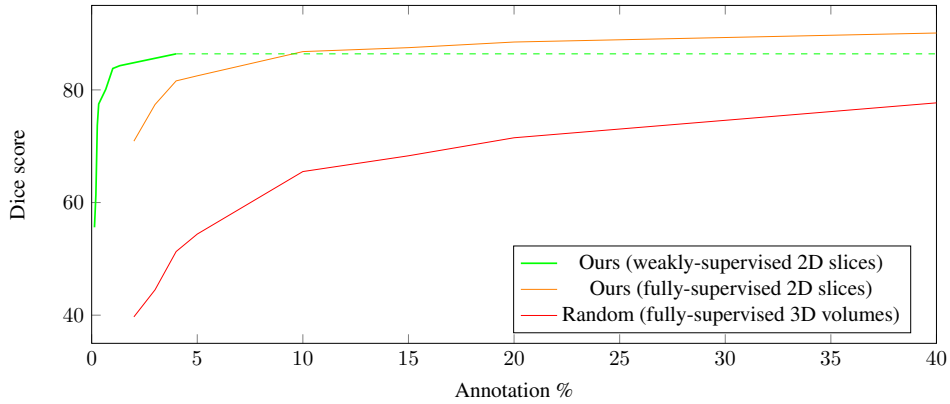


Figure 2: Describes the relationship between model performance and annotation time for our method utilizing weakly and fully-supervised 2D slices and random sampling of fully-supervised 3D volumes on the ACDC dataset. Annotation % is measured as the percentage of the fully-labeled ACDC training data. For weak supervision, we extrapolate the percentage of fully-labeled data based on equivalent annotation time (we follow prior work which assumes that annotators annotate scribbles 15x as fast as the full masks (72)). The dashed green line represents the performance of our method using weakly-supervised 2D slices with 40% of the ACDC training data.

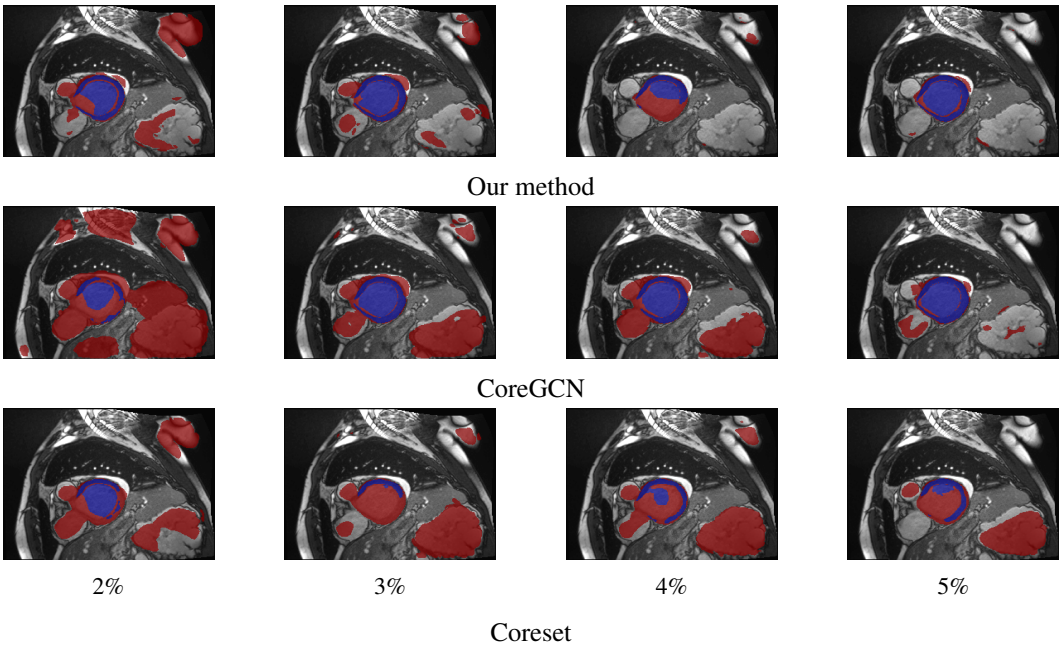


Figure 3: Qualitative comparison of our method, CoreGCN, and Coreset. Blue indicates agreement between model predictions and groundtruth masks and red indicates disagreement.

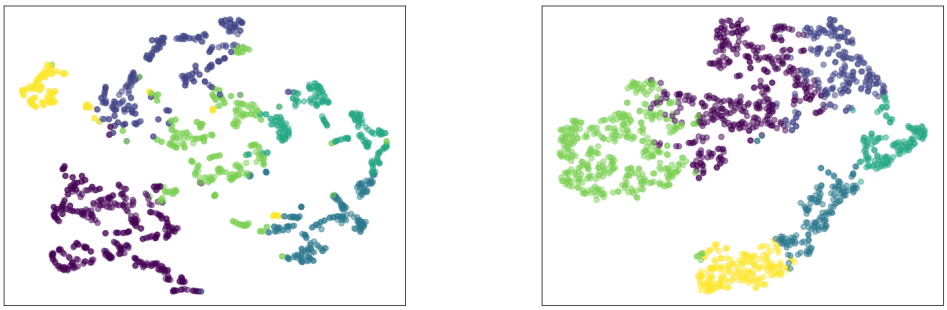
consistently achieving the highest or close to the highest performance for a particular annotation level.

Additionally, our algorithm consistently demonstrates superior performance in both weak and full annotation scenarios, unlike other top-performing methods which struggle in one of these settings. Compared to other algorithms, our method shows superior performance on different datasets as well.

We note that clinically acceptable DICE scores for similar medical segmentation tasks range from 0.5-0.9, depending on the task (68; 25; 75; 8). However, even lower DICE scores can be clinically useful, especially for particular volumes with higher scores or in conjunction with semi-automated segmentation methods (57).

Table 4: Ablation study based on the mean DICE scores for the 2-5% weak annotation datapoint

Coreset	NT-Xent	Patient	Volume	Slice	mDICE
✓					58.5
✓	✓				59.5
✓		✓			<u>60.0</u>
✓			✓		**60.7**
✓				✓	58.1
✓		✓	✓		<u>61.5</u>
✓			✓	✓	60.9
✓	✓		✓		**62.6**
✓	✓	✓	✓		**65.4**
✓		✓	✓	✓	<u>64.1</u>
✓	✓	✓	✓	✓	61.1



(a) NT-Xent Loss (b) Our Loss

Figure 4: t-SNE visualization of dataset clusters generated by different g_ϕ

Our fully-supervised experiments with pre-trained models are in Table 3. We saw improvements with our method in comparison to the best performing comparison methods. Please refer to Appendix D for comprehensive results across all datasets, including calculated bootstrapping standard errors.

4.4 Relationship between model performance and annotation time

In Figure 2 we provided a graph which describes the relationship between model performance and annotation time for our method utilizing weakly and fully-supervised 2D slices and random sampling of fully-supervised 3D volumes on the ACDC dataset. To ensure a fair comparison between the different methods, we do not incorporate any of the results from the pre-trained segmentation models. For the 3D results, similar to the 2D U-Net, we train a 3D U-Net from scratch. Given comparable annotation time, our methods trained on both weakly and fully-supervised 2D slices far exceed the performance of random sampling of 3D volumes and achieve 3D volume maximum performance (with the given budget) with much less annotation time.

4.5 Comparison with related work

Of the diversity-based comparison methods (Coreset, VAAL, TypiClust, Random), the performance for these three are generally worse than our method. For example, our method achieves the best performance on 21 out of 27 comparison points (Tables 1, 2, 3). The next closest is Random sampling, which achieves the best performance on 5 out of 27 comparison points. Of the entropy-based methods (BALD, Variance Ratio, Stochastic Batches), Stochastic Batches achieves the best performance on 4 out of 27 comparison points, the best out of the group.

Coreset and CoreGCN share similarities with our method. However, on almost all the comparison points, they perform much worse. In Figure 3, there is a qualitative comparison of our method, CoreGCN, and Coreset on a difficult slice in the volume after several weak annotation rounds. With more requested annotations, our method is able to reduce errors even in difficult slices. With 5% weak annotation, while Coreset and CoreGCN still retain large error artifacts, our method has minimally visible errors.

4.6 Ablation study

In Table 4, we see our ablation experiments. The first three sections represent our experiments with one, two, and three or more contrastive losses respectively. We note that all the contrastive loss experiments perform better than vanilla Coreset. Additionally, generally the larger combination of losses tend to outperform the smaller combination of losses. The best loss combination involves the volume group loss, patient group loss, and NT-Xent.

One interesting observation is that the loss associated with the volume group — the most diverse group — tends to produce the best additive performance and the loss associated with the the adjacent slice group — the least diverse group — tends to have the worst additive performance. We see this trend in the one loss experiments as well in the two loss and three or more loss experiments, where the combinations with the volume group loss tend to produce the best performance and the combinations with the adjacent slice group loss tends to produce the worst performance.

In order to visualize how effective the learned g_ϕ trained with different losses are for Coreset, we applied k-means clustering to the generated dataset features with a contrastive encoder trained with NT-Xent and trained with our optimal loss and visualized the quality of cluster labels using a t-SNE plot, which can be seen in Figure 4. We note that the clusters from our loss show good cohesion (tightly grouped), separation between clusters, and are easy to differentiate. However, the clusters from NT-Xent show much less cohesion (points are spread over more space) and the separation is less defined. Higher quality clustering emphasizes points are well separated which leads to better performance when trying to find representative points using Coreset.

5 Conclusion

In our research, we introduced a novel metric learning method for Coreset to perform slice-based active learning in 3D medical segmentation. By leveraging diverse data groups in our feature representation, we were able to learn a metric that promoted diversity and our Coreset implementation was able to outperform all existing methods on medical and non-medical datasets in weak and full annotation scenarios with a low annotation budget. Due to limited computational resources, we restricted the number of experiment runs and models we trained. We also acknowledge that we did not fully consider training set bias which can result in unfair outcomes for underrepresented groups. In future work, we hope to remedy some of these issues and focus more robustly on the applications of our approach in diverse domains.

Acknowledgments

This work was partially supported by NSF 2211557, NSF 1937599, NSF 2119643, NSF 2303037, NSF 2312501, NASA, SRC JUMP 2.0 Center, Amazon Research Awards, and Snapchat Gifts.

References

[1] Aziere, N., Todorovic, S.: Ensemble deep manifold similarity learning using hard proxies. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7299–7307 (2019)

[2] Baevski, A., Zhou, Y., Mohamed, A., Auli, M.: wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in neural information processing systems **33**, 12449–12460 (2020)

[3] Beluch, W.H., Genewein, T., Nürnberger, A., Köhler, J.M.: The power of ensembles for active learning in image classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 9368–9377 (2018)

[4] Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.A., Cetin, I., Lekadir, K., Camara, O., Ballester, M.A.G., et al.: Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE transactions on medical imaging **37**(11), 2514–2525 (2018)

[5] Bianconi, F., Fravolini, M.L., Pizzoli, S., Palumbo, I., Minestrini, M., Rondini, M., Nuvoli, S., Spanu, A., Palumbo, B.: Comparative evaluation of conventional and deep learning methods for semi-automated segmentation of pulmonary nodules on ct. Quantitative imaging in medicine and surgery **11**(7), 3286 (2021)

[6] Burmeister, J.M., Rosas, M.F., Hagemann, J., Kordt, J., Blum, J., Shabo, S., Bergner, B., Lippert, C.: Less is more: A comparison of active learning strategies for 3d medical image segmentation. arXiv preprint arXiv:2207.00845 (2022)

[7] Caramalau, R., Bhattarai, B., Kim, T.K.: Sequential graph convolutional network for active learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9583–9592 (2021)

[8] Cardenas, C.E., McCarroll, R.E., Court, L.E., Elgohari, B.A., Elhalawani, H., Fuller, C.D., Kamal, M.J., Meheissen, M.A., Mohamed, A.S., Rao, A., et al.: Deep learning algorithm for auto-delineation of high-risk oropharyngeal clinical target volumes with built-in dice similarity coefficient parameter optimization function. International Journal of Radiation Oncology* Biology* Physics **101**(2), 468–478 (2018)

[9] Chaitanya, K., Erdil, E., Karani, N., Konukoglu, E.: Contrastive learning of global and local features for medical image segmentation with limited annotations. Advances in neural information processing systems **33**, 12546–12558 (2020)

[10] Chen, L., Bentley, P., Mori, K., Misawa, K., Fujiwara, M., Rueckert, D.: Self-supervised learning for medical image analysis using image context restoration. Medical image analysis **58**, 101539 (2019)

[11] Chen, Q., Hong, Y.: Scribble2d5: Weakly-supervised volumetric image segmentation via scribble annotations. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 234–243. Springer (2022)

[12] Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)

[13] Chibane, J., Engelmann, F., Anh Tran, T., Pons-Moll, G.: Box2mask: Weakly supervised 3d semantic instance segmentation using bounding boxes. In: European Conference on Computer Vision. pp. 681–699. Springer (2022)

[14] Chitta, K., Alvarez, J.M., Lesnikowski, A.: Large-scale visual active learning with deep probabilistic ensembles. arXiv preprint arXiv:1811.03575 (2018)

[15] Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: learning dense volumetric segmentation from sparse annotation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19. pp. 424–432. Springer (2016)

[16] Deng, J., Guo, J., Liu, T., Gong, M., Zafeiriou, S.: Sub-center arcface: Boosting face recognition by large-scale noisy web faces. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16. pp. 741–757. Springer (2020)

[17] Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4690–4699 (2019)

[18] Dutta, U.K., Harandi, M., Shekhar, C.C.: Semi-supervised metric learning: A deep resurrection. In: Proceedings of the AAAI Conference on artificial intelligence. vol. 35, pp. 7279–7287 (2021)

[19] Franco, L., Mandica, P., Kallidromitis, K., Guillory, D., Li, Y.T., Galasso, F.: Hyperbolic active learning for semantic segmentation under domain shift. arXiv preprint arXiv:2306.11180 (2023)

[20] Gaillochet, M., Desrosiers, C., Lombaert, H.: Active learning for medical image segmentation with stochastic batches. arXiv preprint arXiv:2301.07670 (2023)

[21] Gal, Y., Islam, R., Ghahramani, Z.: Deep bayesian active learning with image data. In: International conference on machine learning. pp. 1183–1192. PMLR (2017)

[22] Gao, S., Zhou, H., Gao, Y., Zhuang, X.: Bayeseg: Bayesian modeling for medical image segmentation with interpretable generalizability (2023)

[23] Hacohen, G., Dekel, A., Weinshall, D.: Active learning on a budget: Opposite strategies suit high and low budgets. arXiv preprint arXiv:2202.02794 (2022)

[24] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)

[25] Hoebel, K.V.: Domain and User-Centered Machine Learning for Medical Image Analysis. Ph.D. thesis, Massachusetts Institute of Technology (2023)

[26] Huang, Z., Guo, Y., Zhang, N., Huang, X., Decazes, P., Becker, S., Ruan, S.: Multi-scale feature similarity-based weakly supervised lymphoma segmentation in pet/ct images. Computers in Biology and Medicine **151**, 106230 (2022)

[27] Jawade, B., Mohan, D.D., Ali, N.M., Setlur, S., Govindaraju, V.: Napreg: nouns as proxies regularization for semantically aware cross-modal embeddings. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1135–1144 (2023)

[28] Ji, W., Liang, R., Zheng, Z., Zhang, W., Zhang, S., Li, J., Li, M., Chua, T.s.: Are binary annotations sufficient? video moment retrieval via hierarchical uncertainty-based active learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23013–23022 (2023)

[29] Ji, Z., Shen, Y., Ma, C., Gao, M.: Scribble-based hierarchical weakly supervised learning for brain tumor segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part III 22. pp. 175–183. Springer (2019)

[30] Jin, Q., Yuan, M., Qiao, Q., Song, Z.: One-shot active learning for image segmentation via contrastive learning and diversity-based sampling. Knowledge-Based Systems **241**, 108278 (2022)

[31] Ju, J., Jung, H., Oh, Y., Kim, J.: Extending contrastive learning to unsupervised coreset selection. IEEE Access **10**, 7704–7715 (2022)

[32] Jung, S., Kim, S., Lee, J.: A simple yet powerful deep active learning with snapshots ensembles. In: The Eleventh International Conference on Learning Representations (2022)

[33] Kavur, A.E., Selver, M.A., Dicle, O., Barış, M., Gezer, N.S.: CHAOS - Combined (CT-MR) Healthy Abdominal Organ Segmentation Challenge Data (Apr 2019). https://doi.org/10.5281/zenodo.3362844, `https://doi.org/10.5281/zenodo.3362844`

[34] Ke, T.W., Hwang, J.J., Yu, S.X.: Universal weakly supervised segmentation by pixel-to-segment contrastive learning. arXiv preprint arXiv:2105.00957 (2021)

[35] Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. Advances in neural information processing systems **33**, 18661–18673 (2020)

[36] Kim, J., Kim, J., Hwang, S.: Deep active learning with contrastive learning under realistic data pool assumptions. arXiv preprint arXiv:2303.14433 (2023)

[37] Kim, K., Park, D., Kim, K.I., Chun, S.Y.: Task-aware variational adversarial active learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8166–8175 (2021)

[38] Kirsch, A., Van Amersfoort, J., Gal, Y.: Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. Advances in neural information processing systems **32** (2019)

[39] Liu, H., Li, H., Yao, X., Fan, Y., Hu, D., Dawant, B.M., Nath, V., Xu, Z., Oguz, I.: Colossal: A benchmark for cold-start active learning for 3d medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 25–34. Springer (2023)

[40] Luo, X., Hu, M., Liao, W., Zhai, S., Song, T., Wang, G., Zhang, S.: Scribble-supervised medical image segmentation via dual-branch network and dynamically mixed pseudo labels supervision. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 528–538. Springer (2022)

[41] Margatina, K., Vernikos, G., Barrault, L., Aletras, N.: Active learning by acquiring contrastive examples. arXiv preprint arXiv:2109.03764 (2021)

[42] MH Nguyen, D., Nguyen, H., Diep, N., Pham, T.N., Cao, T., Nguyen, B., Swoboda, P., Ho, N., Albarqouni, S., Xie, P., et al.: Lvm-med: Learning large-scale self-supervised vision models for medical imaging via second-order graph matching. Advances in Neural Information Processing Systems **36** (2024)

[43] Michael, O., Georg, W., Horst, P., Horst, B.: Deep metric learning with bier: Boosting independent embeddings robustly. IEEE TPAMI **42**(2), 276–290 (2018)

[44] Militello, C., Rundo, L., Dimarco, M., Orlando, A., Conti, V., Woitek, R., D'Angelo, I., Bartolotta, T.V., Russo, G.: Semi-automated and interactive segmentation of contrast-enhancing masses on breast dce-mri using spatial fuzzy clustering. Biomedical Signal Processing and Control **71**, 103113 (2022)

[45] Mohamadi, S., Amindavar, H.: Deep bayesian active learning, a brief survey on recent advances. arXiv preprint arXiv:2012.08044 (2020)

[46] Mohan, D.D., Jawade, B., Setlur, S., Govindaraju, V.: Deep metric learning for computer vision: A brief overview. Handbook of Statistics **48**, 59–79 (2023)

[47] Movshovitz-Attias, Y., Toshev, A., Leung, T.K., Ioffe, S., Singh, S.: No fuss distance metric learning using proxies. In: Proceedings of the IEEE international conference on computer vision. pp. 360–368 (2017)

[48] Munjal, P., Hayat, N., Hayat, M., Sourati, J., Khan, S.: Towards robust and reproducible active learning using neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 223–232 (2022)

[49] Nath, V., Yang, D., Landman, B.A., Xu, D., Roth, H.R.: Diminishing uncertainty within the training pool: Active learning for medical image segmentation. IEEE Transactions on Medical Imaging **40**(10), 2534–2547 (2020)

[50] Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., Sorkine-Hornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 724–732 (2016)

[51] Pont-Tuset, J., Perazzi, F., Caelles, S., Arbeláez, P., Sorkine-Hornung, A., Van Gool, L.: The 2017 davis challenge on video object segmentation. arXiv preprint arXiv:1704.00675 (2017)

[52] Pop, R., Fulop, P.: Deep ensemble bayesian active learning: Addressing the mode collapse issue in monte carlo dropout via ensembles. arXiv preprint arXiv:1811.03897 (2018)

[53] Qian, Q., Shang, L., Sun, B., Hu, J., Li, H., Jin, R.: Softtriple loss: Deep metric learning without triplet sampling. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6450–6458 (2019)

[54] Qian, Z., Li, K., Lai, M., Chang, E.I.C., Wei, B., Fan, Y., Xu, Y.: Transformer based multiple instance learning for weakly supervised histopathology image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 160–170. Springer (2022)

[55] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)

[56] Ren, H., Zhou, L., Liu, G., Peng, X., Shi, W., Xu, H., Shan, F., Liu, L.: An unsupervised semi-automated pulmonary nodule segmentation method based on enhanced region growing. Quantitative Imaging in Medicine and Surgery **10**(1), 233 (2020)

[57] Rhee, D.J., Cardenas, C.E., Elhalawani, H., McCarroll, R., Zhang, L., Yang, J., Garden, A.S., Peterson, C.B., Beadle, B.M., Court, L.E.: Automatic detection of contouring errors using convolutional neural networks. Medical physics **46**(11), 5086–5097 (2019)

[58] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015)

[59] Rossetti, S., Zappia, D., Sanzari, M., Schaerf, M., Pirri, F.: Max pooling with vision transformers reconciles class and shape in weakly supervised semantic segmentation. In: European Conference on Computer Vision. pp. 446–463. Springer (2022)

[60] Roth, K., Brattoli, B., Ommer, B.: Mic: Mining interclass characteristics for improved metric learning. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8000–8009 (2019)

[61] Roth, K., Vinyals, O., Akata, Z.: Integrating language guidance into vision-based deep metric learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16177–16189 (2022)

[62] Sakinis, T., Milletari, F., Roth, H., Korfiatis, P., Kostandy, P., Philbrick, K., Akkus, Z., Xu, Z., Xu, D., Erickson, B.J.: Interactive segmentation of medical images through fully convolutional neural networks. arXiv preprint arXiv:1903.08205 (2019)

[63] Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 815–823 (2015)

[64] Sener, O., Savarese, S.: Active learning for convolutional neural networks: A core-set approach. arXiv preprint arXiv:1708.00489 (2017)

[65] Siddhant, A., Lipton, Z.C.: Deep bayesian active learning for natural language processing: Results of a large-scale empirical study. arXiv preprint arXiv:1808.05697 (2018)

[66] Sinha, S., Ebrahimi, S., Darrell, T.: Variational adversarial active learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5972–5981 (2019)

[67] Sohn, K.: Improved deep metric learning with multi-class n-pair loss objective. Advances in neural information processing systems **29** (2016)

[68] Sun, Y., Wang, Y., Gan, K., Wang, Y., Chen, Y., Ge, Y., Yuan, J., Xu, H.: Reliable delineation of clinical target volumes for cervical cancer radiotherapy on ct/mr dual-modality images. Journal of Imaging Informatics in Medicine pp. 1–14 (2024)

[69] Taleb, A., Loetzsch, W., Danz, N., Severin, J., Gaertner, T., Bergner, B., Lippert, C.: 3d self-supervised methods for medical imaging. Advances in neural information processing systems **33**, 18158–18172 (2020)

[70] Teh, E.W., DeVries, T., Taylor, G.W.: Proxynca++: Revisiting and revitalizing proxy neighborhood component analysis. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16. pp. 448–464. Springer (2020)

[71] Tsai, D.M., Fan, S.K.S., Chou, Y.H.: Auto-annotated deep segmentation for surface defect detection. IEEE Transactions on Instrumentation and Measurement **70**, 1–10 (2021)

[72] Valvano, G., Leo, A., Tsaftaris, S.A.: Learning to segment from scribbles using multi-scale adversarial attention gates. IEEE Transactions on Medical Imaging **40**(8), 1990–2001 (2021)

[73] Vepa, A., Choi, A., Nakhaei, N., Lee, W., Stier, N., Vu, A., Jenkins, G., Yang, X., Shergill, M., Desphy, M., et al.: Weakly-supervised convolutional neural networks for vessel segmentation in cerebral angiography. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 585–594 (2022)

[74] Wang, C., Zheng, W., Li, J., Zhou, J., Lu, J.: Deep factorized metric learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7672–7682 (2023)

[75] Wang, S., Li, C., Wang, R., Liu, Z., Wang, M., Tan, H., Wu, Y., Liu, X., Sun, H., Yang, R., et al.: Annotation-efficient deep learning for automatic medical image segmentation. Nature communications **12**(1), 5915 (2021)

[76] Wang, T., Li, X., Yang, P., Hu, G., Zeng, X., Huang, S., Xu, C.Z., Xu, M.: Boosting active learning via improving test performance. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 8566–8574 (2022)

[77] Wen, Y., Zhang, K., Li, Z., Qiao, Y.: A discriminative feature learning approach for deep face recognition. In: Computer vision–ECCV 2016: 14th European conference, amsterdam, the netherlands, October 11–14, 2016, proceedings, part VII 14. pp. 499–515. Springer (2016)

[78] Wu, F., Zhuang, X.: Minimizing estimated risks on unlabeled data: A new formulation for semi-supervised medical image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(5), 6021–6036 (2022)

[79] Xie, B., Yuan, L., Li, S., Liu, C.H., Cheng, X., Wang, G.: Active learning for domain adaptation: An energy-based approach. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 8708–8716 (2022)

[80] Xuan, H., Souvenir, R., Pless, R.: Deep randomized ensembles for metric learning. In: Proceedings of the European conference on computer vision (ECCV). pp. 723–734 (2018)

[81] Yang, L., Zhang, Y., Chen, J., Zhang, S., Chen, D.Z.: Suggestive annotation: A deep active learning framework for biomedical image segmentation. In: Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part III 20. pp. 399–407. Springer (2017)

[82] Zhang, B., Xiao, J., Jiao, J., Wei, Y., Zhao, Y.: Affinity attention graph neural network for weakly supervised semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(11), 8082–8096 (2021)

[83] Zhang, H., Burrows, L., Meng, Y., Sculthorpe, D., Mukherjee, A., Coupland, S.E., Chen, K., Zheng, Y.: Weakly supervised segmentation with point annotations for histopathology images via contrast-based variational model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15630–15640 (2023)

[84] Zhang, K., Zhuang, X.: Cyclemix: A holistic strategy for medical image segmentation from scribble supervision. arXiv preprint arXiv:2203.01475 (2022)

[85] Zhang, Y., Zhang, X., Xie, L., Li, J., Qiu, R.C., Hu, H., Tian, Q.: One-bit active query with contrastive pairs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9697–9705 (2022)

[86] Zheng, H., Han, J., Wang, H., Yang, L., Zhao, Z., Wang, C., Chen, D.Z.: Hierarchical self-supervised learning for medical image segmentation based on multi-domain data aggregation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24. pp. 622–632. Springer (2021)

[87] Zhu, Y., Xu, W., Liu, Q., Wu, S.: When contrastive learning meets active learning: A novel graph active learning paradigm with self-supervision. arXiv preprint arXiv:2010.16091 (2020)

[88] Zhuang, F., Moulin, P.: Deep semi-supervised metric learning with mixed label propagation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3429–3438 (2023)

[89] Zhuang, X.: Multivariate mixture model for myocardial segmentation combining multi-source images. IEEE transactions on pattern analysis and machine intelligence **41**(12), 2933–2946 (2018)

A Loss weights

In our experiments and ablation study, a 1.0 weight was applied whenever the NT-Xent loss was used. Additionally, a 1.0 weight was applied if only one group contrastive loss was used. In the ablation study on the ACDC dataset, for the experiments with multiple group contrastive losses, we tried different combinations of weights for the group contrastive losses and reported the best results. We will utilize the formulation from Equation 5 for the loss weights and report them as a four-tuple (a, b, c, d) which corresponds to $(\lambda_0, \lambda_1, \lambda_2, \lambda_3)$ (in which λ_0 is 1 if the NT-Xent loss was used and 0 if it is not). The combinations we tried are as follow (with the best bolded):

- patient and volume group loss without NT-Xent loss: (0, 0.125, 0.875, 0), **(0, 0.50, 0.50, 0)**
- volume and slice group loss without NT-Xent loss: (0, 0, 0.125, 0.875), **(0, 0, 0.50, 0.50)**
- patient and volume group loss with NT-Xent loss: (1,0. 10 , 0.35, 0), (1, 0.117, 0.233, 0), **(1, 0.05, 0.35, 0)**
- patient, volume, and slice group loss without NT-Xent loss: (0, 0.05, 0.25, 0.7), **(0, 0.33, 0.33, 0.33)**
- patient, volume, and slice group loss with NT-Xent loss: (1, 0.10, 0.20, 0.05), **(1, 0.05, 0.35, 0.025)**

We utilized the best loss/weight combination for our ACDC experiments. For the CHAOS, MSCMR, and DAVIS datasets, because there is only one volume per patient and thus no difference between the patient and volume loss, we tested two loss/weight combinations:

- volume loss with weight 0.35 with NT-Xent loss
- volume loss with weight 0.10 and slice loss with weight 0.30 with NT-Xent loss

Both weight combinations performed better than other comparison methods though volume loss with weight 0.35 with NT-Xent loss performed slightly better than the other combination.

For the pre-trained weights, we found that best results were obtained on the ACDC dataset with just the patient group without the NT-Xent loss. We tested different combinations of groups but found the results were worse. We used the same weight setting for CHAOS and DAVIS

B Theoretical bounds for loss

First, we assume that the expectation over the data distribution of the volume-based loss and slice-based loss are equivalent. Formally, this is described as follows:

$$E_{\mathbf{x},\mathbf{y}\sim p_{\mathcal{Z}}}[\mathcal{L}_{volume}(\hat{\mathbf{y}}, \mathbf{y})] = E_{\mathbf{x}',\mathbf{y}'\sim p_{\mathcal{Z}'}}[\mathcal{L}_{slice}(\hat{\mathbf{y}}', \mathbf{y}')]$$

where $\mathcal{L}_{volume} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ and $\mathcal{L}_{slice} : \mathcal{X}' \times \mathcal{Y}' \to \mathbb{R}$ are the volume-based and slice-based loss respectively. For the rest of the proof, $\mathcal{L}_{slice}(\cdot)$ is referred to by $\mathcal{L}(\cdot)$. Following the derivation provided in the original Coreset paper (64) we have:

$$\begin{aligned}
E_{\mathbf{x}',\mathbf{y}'\sim p_{\mathcal{Z}'}}[\mathcal{L}(\hat{\mathbf{y}}', \mathbf{y}')] &\leq \underbrace{\left| E_{\mathbf{x}',\mathbf{y}'\sim p_{\mathcal{Z}'}}[\mathcal{L}(\hat{\mathbf{y}}', \mathbf{y}')] - \frac{1}{n}\sum_{i\in[n]} \mathcal{L}(\hat{\mathbf{y}}', \mathbf{y}') \right|}_{\text{Generalization Error}} + \underbrace{\frac{1}{|\mathbf{s}|}\sum_{j\in\mathbf{s}} \mathcal{L}(\hat{\mathbf{y}}', \mathbf{y}')}_{\text{Training Error}} \\
&+ \underbrace{\left| \frac{1}{n}\sum_{i\in[n]} \mathcal{L}(\hat{\mathbf{y}}', \mathbf{y}') - \frac{1}{|\mathbf{s}|}\sum_{j\in\mathbf{s}} \mathcal{L}(\hat{\mathbf{y}}', \mathbf{y}'), \right|}_{\text{Core-Set Loss}} \\
&\leq \underbrace{\left| \frac{1}{n}\sum_{i\in[n]} \mathcal{L}(\hat{\mathbf{y}}', \mathbf{y}') - \frac{1}{|\mathbf{s}|}\sum_{j\in\mathbf{s}} \mathcal{L}(\hat{\mathbf{y}}', \mathbf{y}'), \right|}_{\text{Core-Set Loss}}
\end{aligned}$$

We get the last line of the inequality because we assume the Training Error is zero and, similar to the original Coreset paper, we assume the Generalization Error is zero, (which is a reasonable assumption because most CNNs have very small generalization error) (64).

Thus, our active learning objective can be re-defined as:

$$\underset{s^1 \subset D, |s^1| \leq k}{\operatorname{argmin}} \frac{1}{n} \sum_{i \in [n]} \mathcal{L}(\hat{\mathbf{y}}', \mathbf{y}') - \frac{1}{|\mathbf{s}|} \sum_{j \in \mathbf{s}} \mathcal{L}(\hat{\mathbf{y}}', \mathbf{y}') \tag{6}$$

We will now present the following theorem:

Theorem 1. *Given n i.i.d. samples drawn from $p_{\mathcal{Z}'}$ as $\{\mathbf{x}'_i, \mathbf{y}'_i\}_{i \in [n]}$, and set of points $\mathbf{s}$. If loss function $\mathcal{L}(\hat{\mathbf{y}}', \mathbf{y}')$ is λ^l-Lipschitz continuous for all $\hat{\mathbf{y}}', \mathbf{y}'$ and bounded by L, segmentation function $\eta_{\mathbf{c}}(\mathbf{x}') = p(\mathbf{y}' = \mathbf{c}|\mathbf{x}')$ is λ^η-Lipshitz continuous for all $\mathbf{x}' \in \Omega'$ and $\mathbf{c} \in \mathcal{Y}'$, $\mathbf{s}$ is δ cover of $\{\mathbf{x}'_i, \mathbf{y}'_i\}_{i \in [n]}$, and $l(\hat{\mathbf{y}}'_{s(j)}, \mathbf{y}'_{s(j)}) = 0 \quad \forall j \in [m]$; with probability at least $1 - \gamma$,*

$$\frac{1}{n} \sum_{i \in [n]} \mathcal{L}(\hat{\mathbf{y}}'_i, \mathbf{y}'_i) - \frac{1}{|\mathbf{s}|} \sum_{j \in \mathbf{s}} \mathcal{L}(\hat{\mathbf{y}}'_j, \mathbf{y}'_j) \leq \delta(\lambda^l + \lambda^\eta L 2^{hwdk}) + \sqrt{\frac{L^2 \log(1/\gamma)}{2n}}.$$

Similar to (64), we can start our proof with bounding $E_{\mathbf{y}'_i \sim \eta(\mathbf{x}'_i)} \mathcal{L}(\hat{\mathbf{y}}'_i, \mathbf{y}'_i)$. Following a similar approach to (64), we get:

$$E_{\mathbf{y}'_i \sim \eta(\mathbf{x}'_i)} \mathcal{L}(\hat{\mathbf{y}}'_i, \mathbf{y}'_i) \leq \delta(\lambda^l + \lambda^\eta L 2^{hwdk})$$

Again, following (64), we can use the Hoeffding's Bound to conclude the rest of the proof.

C Batch sampler

The pseudocode for the batch sampler can be seen in Algorithm 1. This implementation assumes the use of all three group contrastive losses. If the adjacent slice group is removed, then remove line 6. If the volume group is removed, then remove line 7. If the patient group is removed, then remove line 8. An illustration on how the batch sampler works can be seen in Figure 5

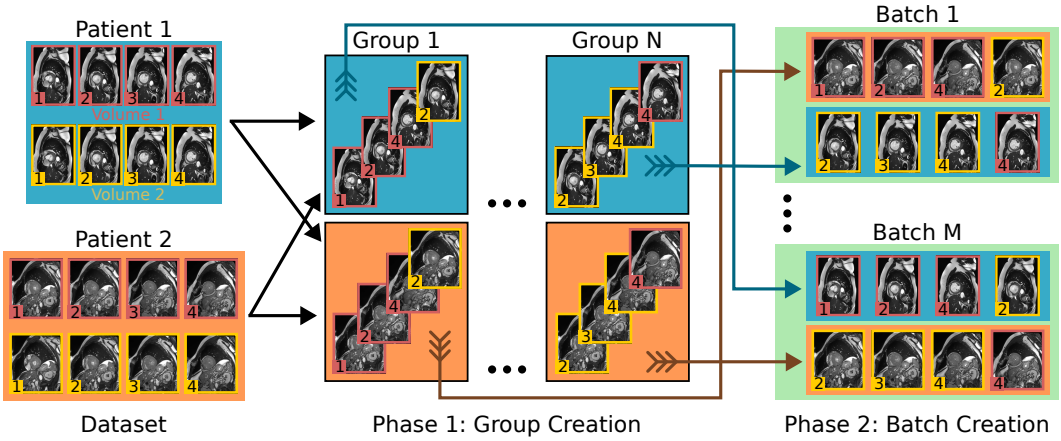


Figure 5: Overview of the batch sampler for Group-based Contrastive Learning

Algorithm 1 Batch sampler for one epoch

Input: List of patient groups $P_1, P_2, ..., P_L$. Each patient group i contains a list of volume groups $V_1^i, V_2^i, ..., V_K^i$. Each volume group j for patient i contains a list of slices $s_1^{ij}, s_2^{ij}, ..., s_T^{ij}$. Batch size is M.

$PG \leftarrow list()$
for all $P_1, P_2, ..., P_L$ **do**
  $groups \leftarrow ()$
  for all $V_1^i, V_2^i, ..., V_K^i$ **do**
    for all $s_1^{ij}, s_2^{ij}, ..., s_T^{ij}$ **do**
      $s \leftarrow RandomChoice(s_{k-1}^{ij}, s_{k+1}^{ij})$
      $v \leftarrow RandomChoice(\cup_{t \neq k} s_t^{ij})$
      $p \leftarrow RandomChoice(\cup_{V \in P_i : s \in V} s / \{s_j^{ij}\})$
      $group \leftarrow [s_k^{ij}, s, v, p]$
      Add $group$ to $groups$
    end for
  end for
  Add $groups$ to PG
end for
$batches \leftarrow ()$
while $|PG| \geq M$ **do**
  $batch \leftarrow list()$
  $S \leftarrow ()$
  while $|batch| < M$ **do**
    $groups \leftarrow RandomChoice(\cup_{g \in PG \wedge g \notin S})$
    $group \leftarrow RandomChoice(groups)$
    $groups \leftarrow groups \backslash group$
    if $|groups| = 0$ **then**
      $PG \leftarrow PG \backslash groups$
    else
      Add $groups$ to S
    end if
    Extend $batch$ with $group$
  end while
  Add $batch$ to $batches$
end while
return batches

D Additional results

We have included all collected results on all the datasets. We also report twice the bootstrap standard error for the means over the different experiment runs. Our process is as follows. We generate 1000 bootstraps. For each bootstrap, we generate a resampling with replacement of the test set volumes and generate the mean of the test set metrics (average 3D Dice score over the test set) over the different experiment runs at the desired evaluation point. We then calculate the sample standard deviation of the bootstrapped means and multiply by two.

Because the CHAOS and DAVIS have fewer volumes (and the bootstrapping errors are artifically inflated), rather generating a resampling with replacement on the test set volumes, we generate a resampling over the slices.

Table 5: DICE scores for ACDC

	Weakly-supervised			
	2%	3%	4%	5%
BALD	44.8±1.9	54.8±1.9	61.4±1.8	66.2±1.6
Variance Ratio	43.3±2.1	45.0±2.2	54.2±2.4	61.4±2.1
Random	44.9±2.3	45.7±2.2	59.2±2.1	66.9±1.9
VAAL	41.8±2.3	47.8±2.2	66.1±2.1	72.1±1.9
Coreset	45.2±2.1	49.8±2.0	68.7±2.0	70.1±1.8
TypiClust	44.8±2.2	45.6±2.0	67.6±1.8	73.1±1.6
Stochastic Batches	36.9±2.2	39.5±2.0	53.7±1.8	57.4±1.5
CoreGCN	40.8±2.2	49.3±2.0	69.1±1.9	74.1±1.7
Ours	55.6±2.1	61.4±1.9	73.7±2.2	77.5±1.6
	10%	15%	20%	40%
BALD	77.7±1.3	81.6±1.2	83.3±1.1	86.4±0.9
Variance Ratio	74.3±1.9	83.2±1.5	84.4±1.2	85.9±0.9
Random	76.5±1.4	83.0±1.3	84.1±1.1	86.7±0.9
VAAL	79.6±1.4	82.4±1.2	84.4±1.1	86.4±0.9
Coreset	80.3±1.3	83.6±1.2	85.3±1.1	86.9±1.0
TypiClust	77.7±1.4	82.0±1.3	83.6±1.3	85.7±1.2
Stochastic Batches	69.2±1.4	78.5±1.3	81.9±1.2	85.8±1.1
CoreGCN	78.4±1.3	82.9±1.3	83.9±1.1	86.3±0.9
Ours	80.1±1.4	83.8±1.3	84.3±1.2	86.4±1.0
	Fully-supervised			
	2%	3%	4%	5%
BALD	53.8±3.7	66.0±2.7	67.9±2.5	71.7±2.2
Variance Ratio	52.6±3.3	62.2±3.2	66.6±3.6	69.1±3.3
Random	66.3±3.2	76.9±3.0	79.3±2.4	80.1±1.8
VAAL	63.2±2.6	75.2±2.3	79.5±2.1	81.3±1.8
Coreset	58.9±3.7	69.3±3.3	75.7±3.6	81.9±2.7
TypiClust	66.6±3.4	75.8±3.1	79.5±2.4	81.7±2.1
Stochastic Batches	60.3±3.0	69.8±2.8	74.1±2.6	75.4±2.5
CoreGCN	67.7±2.6	74.9±2.4	79.0±2.2	80.7±2.1
Ours	70.9±2.7	77.4±2.4	81.6±2.2	82.5±2.1
	10%	15%	20%	40%
BALD	81.6±1.8	85.5±1.2	85.9±1.0	89.3±0.9
Variance Ratio	76.4±3.2	80.7±3.0	83.6±2.4	86.9±1.3
Random	86.8±1.6	88.1±1.3	88.3±1.1	90.2±0.9
VAAL	85.9±1.4	88.0±1.3	87.9±1.2	89.9±1.1
Coreset	85.8±2.2	87.7±1.3	88.8±1.1	90.2±1.0
TypiClust	85.9±1.3	87.1±1.2	88.0±1.0	89.5±0.9
Stochastic Batches	81.6±2.4	84.6±1.4	86.5±1.3	89.3±1.1
CoreGCN	85.9±1.3	87.9±1.2	88.5±1.2	89.6±1.0
Ours	86.8±1.8	87.5±1.4	88.5±1.3	90.2±1.0

Table 6: DICE scores for MSCMR

	Weakly-supervised			
	2%	3%	4%	5%
BALD	37.0±2.8	48.9±2.9	57.4±2.7	60.0±3.0
Variance Ratio	41.3±5.0	46.8±5.1	52.8±5.5	55.3±3.8
Random	39.7±3.7	55.0±4.0	61.0±4.1	61.5±3.1
Coreset	28.7±5.0	53.6±4.5	56.9±4.1	58.7±3.6
Stochastic Batches	38.5±5.6	56.2±4.2	59.8±4.6	60.5±3.6
CoreGCN	27.0±5.6	48.7±5.3	57.1±5.2	57.2±3.6
Ours	44.3±5.2	53.3±4.9	63.4±5.4	63.5±4.5
	10%	15%	20%	40%
BALD	77.0±2.8	79.5±3.2	80.1±3.0	85.8±2.3
Variance Ratio	65.8±3.3	72.9±3.1	78.4±2.9	84.3±2.6
Random	76.0±2.9	82.4±3.1	83.2±3.0	85.7±2.8
Coreset	74.4±2.8	81.4±2.1	83.1±2.5	86.7±1.9
Stochastic Batches	76.1±3.2	80.6±2.8	83.1±2.2	86.2±1.9
CoreGCN	71.0±3.6	80.9±3.7	83.0±3.8	85.5±3.8
Ours	77.2±2.9	82.4±2.3	83.6±2.1	86.3±2.3

Table 7: DICE scores for CHAOS

	Fully-supervised			
	2%	3%	4%	5%
BALD	79.6±1.8	80.7±1.8	80.4±1.8	81.1±1.9
Variance Ratio	74.9±2.5	72.9±2.7	76.6±2.6	76.2±2.6
Random	80.7±1.9	81.7±1.7	84.2±1.5	85.1±1.3
Coreset	80.0±1.8	80.4±2.0	81.2±1.9	88.1±1.0
Stochastic Batches	77.2±2.3	82.9±1.9	83.8±1.8	84.7±2.0
CoreGCN	67.8±2.1	77.7±2.1	77.8±1.2	74.4±1.5
Ours	80.5±1.8	82.5±1.5	85.9±0.8	90.3±1.3
	10%	15%	20%	40%
BALD	92.6±0.6	94.5±0.5	94.9±0.5	95.8±0.4
Variance Ratio	82.6±2.4	85.1±2.0	87.2±1.9	92.8±1.2
Random	92.7±0.8	94.3±0.6	94.8±0.6	96.2±0.3
Coreset	92.2±0.9	94.9±0.5	95.8±0.4	96.5±0.2
Stochastic Batches	92.1±1.1	93.0±1.0	94.1±0.7	96.1±0.3
CoreGCN	85.9±1.0	93.3±0.6	94.1±0.5	94.3±0.3
Ours	92.5±0.7	94.4±0.7	95.6±0.5	96.3±0.3

Table 8: DICE scores for DAVIS

	Fully-supervised			
	10%	20%	30%	40%
BALD	43.6±0.6	42.0±0.5	43.4±0.4	43.8±0.6
Variance Ratio	36.1±0.5	31.2±0.4	34.9±0.6	40.7±0.3
Random	39.6±0.6	40.5±0.7	47.4±0.7	48.5±0.5
Coreset	31.7±0.7	39.4±0.6	42.2±0.6	42.1±0.5
Stochastic Batches	40.3±0.6	41.5±0.8	45.1±0.6	47.6±0.6
Ours	42.8±0.7	45.2±0.6	45.5±0.7	46.6±0.7

Table 9: DICE scores for ACDC (pretrained)

	Fully-supervised				
	1%	2%	3%	4%	5%
Random	61.9±1.8	77.8±1.4	81.9±1.2	85.0±0.9	85.6±0.9
Stochastic Batches	55.2±1.7	79.3±1.3	82.5±1.2	83.4±1.0	86.4±1.0
Coreset	64.0±1.5	78.3±1.4	81.7±1.3	83.7±1.1	84.0±1.2
Ours	66.1±1.9	79.4±1.4	82.0±1.3	84.3±1.0	85.9±1.0

Table 10: DICE scores for CHAOS (pretrained)

	Fully-supervised				
	1%	2%	3%	4%	5%
Random	91.8±0.5	94.6±0.4	95.9±0.3	96.2±0.2	96.4±0.2
Stochastic Batches	92.4±0.5	94.5±0.3	95.5±0.3	95.9±0.3	96.3±0.3
Coreset	92.3±0.8	94.8±0.4	95.7±0.2	96.3±0.2	96.4±0.2
Ours	93.4±0.3	95.1±0.3	95.4±0.2	96.1±0.2	96.1±0.2

Table 11: DICE scores for DAVIS (pretrained)

	Fully-supervised				
	1%	2%	3%	4%	5%
Random	71.0±0.5	71.2±0.5	76.7±0.6	76.2±0.7	79.3±0.7
Stochastic Batches	69.1±0.6	73.5±0.7	76.9±0.7	76.8±0.7	78.9±0.6
Coreset	68.5±0.6	73.2±0.6	76.9±0.6	76.9±0.7	77.4±0.6
Ours	71.1±0.6	73.5±0.7	76.4±0.6	77.2±0.7	77.2±0.5

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes they do.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Experimental limitations are discussed in Section 5. Computational requirements are discussed in Section 4.2.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We provide a full set of assumptions and complete proof in Appendix B.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All implementation and experiment details are discussed in Section 3.2.1, **??**, and 4.2 and Appendix. We have also provided our code in the supplementary materials and the data is publicly available. We will also post our code to Github.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We have provided code and links to data in the supplementary materials. The code will also be posted to Github.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Yes, this information is provided in Section 3.2.1 and 4.2, Appendix, and can also be viewed in the code in the supplementary materials or on Github.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Yes, we report bootstrapping standard errors for our results and describe our approach in Appendix D.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   We discuss this in Section 4.2.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Because we used existing datasets with human subjects, in Section 4.1 we discussed whether the dataset owners followed data collection regulations and whether human consent was obtained.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: While positive social impacts are discussed throughout the paper (especially in the introduction), we also briefly discussed that we did not consider issues of bias which can have unfair outcomes for underrepresented groups in Section 5.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our models do not have a high risk of misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [No]

    Justification: For the datasets we used, no license information was available and we were unable to reach out to the asset creators. However, we have properly credited the original owners of all the datasets and followed all the owners' directions for research use of the datasets. The license and terms of use for the segmentation model repository are both followed and provided within the wsl4mis subdirectory in the code submitted in the supplementary materials.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.

- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The only assets we are providing is our code, which is provided in the supplementary materials and on Github. The license information as well as documentation on how to use the code is provided within the code directory. Other relevant details can also be referenced in the paper.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [No]

    Justification: We used existing datasets that contain human subject data. As mentioned before, in Section 4.1 we discussed whether the dataset owners followed data collection regulations and whether human consent was obtained.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [No]

Justification: We used existing datasets that contain human subject data. While we were not able to obtain all of this information, as mentioned before in Section 4.1 we discussed whether the dataset owners followed data collection regulations and whether human consent was obtained.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.