Medical Image Analysis (2023)

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Active learning for medical image segmentation with stochastic batches

Mélanie Gaillochet[a,*], Christian Desrosiers[a], Hervé Lombaert[a]

[a]*ETS Montréal, 1100 Notre-Dame St W, Montreal H3C 1K3, QC, Canada*

ARTICLE INFO

Article history:
Received September 15, 2023

Keywords: Active learning, Segmentation, Medical image analysis, Uncertainty

ABSTRACT

The performance of learning-based algorithms improves with the amount of labelled data used for training. Yet, manually annotating data is particularly difficult for medical image segmentation tasks because of the limited expert availability and intensive manual effort required. To reduce manual labelling, active learning (AL) targets the most informative samples from the unlabelled set to annotate and add to the labelled training set. On the one hand, most active learning works have focused on the classification or limited segmentation of natural images, despite active learning being highly desirable in the difficult task of medical image segmentation. On the other hand, uncertainty-based AL approaches notoriously offer sub-optimal batch-query strategies, while diversity-based methods tend to be computationally expensive. Over and above methodological hurdles, random sampling has proven an extremely difficult baseline to outperform when varying learning and sampling conditions. This work aims to take advantage of the diversity and speed offered by random sampling to improve the selection of uncertainty-based AL methods for segmenting medical images. More specifically, we propose to compute uncertainty at the level of batches instead of samples through an original use of stochastic batches (SB) during sampling in AL. Stochastic batch querying is a simple and effective add-on that can be used on top of any uncertainty-based metric. Extensive experiments on two medical image segmentation datasets show that our strategy consistently improves conventional uncertainty-based sampling methods. Our method can hence act as a strong baseline for medical image segmentation. The code is available on: https://github.com/Minimel/StochasticBatchAL.git.

1. Introduction

Data annotation is fundamental to medical imaging. Notably, the performance of segmentation algorithms depends on the amount of annotated training data. The manual annotation of pixel-level ground truth is therefore highly sought but remains difficult to obtain due to two challenging problems. First, the pixel-wise annotation of entire biological structures is a laborious and expensive task that requires highly trained clinicians. Second, image acquisition grows faster than the experts' ability to manually process the data, leaving large datasets mostly unlabelled. Clinicians can realistically annotate only small sets of images with a limited capacity to scale up. This constraint creates a need for strategies that reduce the crucial but arduous annotation efforts in medical imaging.

To maximize the performance of a model with reduced annotated data during training, two types of approaches can unleash the potential of unlabelled data: active learning and semi-supervised learning. Active learning (AL) aims to identify the best samples to annotate and use during training. Meanwhile, semi-supervised learning seeks to improve the representation

*Corresponding author: email: melanie.gaillochet.1@ens.etsmtl.ca

learned from data by exploiting unlabelled samples in addition to the few labelled ones. However, this approach still leaves the question of choosing which samples to use for the labelled set, underlining the importance of active learning.

Images in the training set do not contribute equally to the performance of learning-based algorithms (Settles, 2009). Given a large unlabelled dataset, active learning overcomes labelled data scarcity by incrementally identifying the most valuable samples to be annotated and added to a training set (Budd et al., 2021; Ren et al., 2021). Actively selecting which data to label conceivably maximizes the performance of machine learning models with a minimum amount of labelled data. AL strategies also have the potential of accelerating training convergence and improving robustness by targeting specific types of data points (Nath et al., 2021).

Active learning methods can be divided into three broad categories: uncertainty-based sampling strategies, representative-based sampling strategies and hybrid approaches (Settles, 2009; Budd et al., 2021). Uncertainty-based methods assume that the most valuable samples to annotate are the ones for which the current model is least confident. These methods, which differ in ways of calculating uncertainty, are however susceptible to target outlier samples or redundant information, particularly when querying batches of samples. To avoid bias towards narrow locals in distributions, representative-based and hybrid approaches try to diversify the set of candidate samples. Ensuring such diversity generally relies on learning a latent data representation, which requires estimating pairwise distances between all samples or computing their marginal distribution. These strategies consequently hardly scale satisfyingly to high dimensions. Consequently, the majority of active learning approaches applied to computer vision focus on lower-dimensional tasks such as classification, while AL approaches for segmentation tend to focus on natural images with several thousands of annotated images (Sinha et al., 2019; Huang et al., 2021; Kim et al., 2021; Xie et al., 2022). Due to its high-dimensional nature, medical image segmentation remains an ongoing challenge in active learning, despite the substantial need to minimize the high cost of manual annotation from clinical expertise.

A limited yet increasing number of works acknowledges that random sampling is, in practice, a painstakingly difficult baseline to outperform in active learning (Kirsch et al., 2019; Mittal et al., 2019; Nath et al., 2021; Munjal et al., 2022; Burmeister et al., 2022). Indeed, the gains of AL strategies over random sampling are often inconsistent across different experimental setups. For example, varying the sampling budget can cancel the improvements originally observed for such strategies (Bengar et al., 2021; Munjal et al., 2022). Similarly, existing methods for AL tend to be sensitive to the model architecture, hyperparameters and regularization used during training (Mittal et al., 2019; Munjal et al., 2022). These hurdles hinder AL advances in medical image segmentation.

This paper intends to address the limitations of current AL methods, notably their drawback of selecting batches solely based on per-sample uncertainty, the computational cost of ensuring diversity, and the significantly varying amounts of robustness in performance across experimental setups. Our work proposes to leverage the power of randomness during uncertainty-based batch sampling to improve the overall segmentation performance of AL models.

1.1. Contributions

We introduce the use of stochastic batch (SB) querying, a simple and effective add-on to uncertainty-based AL strategies, compatible with any uncertainty metric (see Fig.1). Our stochastic batch sampling strategy proves advantageous by:

1. minimizing the problem of uncertainty-based strategies, often susceptible to query samples with redundant information;
2. allowing uncertainty-based AL strategies to benefit from a larger diversity of samples in a simple and computationally-efficient way; and
3. providing noticeably consistent gains across different experimental settings, as shown by our extensive ablation studies.

2. Literature review

Active learning methods maximize the future model performance by augmenting the current labelled training set with the most informative unlabelled samples. AL approaches mainly fall into uncertainty-based, representative-based or hybrid strategies, each described next.

2.1. Uncertainty-based AL methods

Uncertainty is one of the most prevalent criteria for sampling in active learning. Uncertainty-based methods query samples for which the current model is least confident (Settles, 2009). AL strategies for deep learning-based models have initially applied traditional AL methods that identify difficult examples using simple heuristics. However, in practice, they still hardly scale to high-dimensional data (Beluch et al., 2018) or are not consistently effective for deep learning models that rely on batch selection (Sener and Savarese, 2018; Ren et al., 2021). Hence, subsequent work has combined traditional uncertainty measures, such as the entropy of the output probabilities, with measures of geometric uncertainty (Konyushkova et al., 2019) or with the pseudo-labelling of samples with confident predictions (Wang et al., 2017). Similarly, Gal et al. (2017) and Kirsch et al. (2019) adapt existing heuristics to a Bayesian framework through Monte Carlo dropout. More recently, Yoo and Kweon (2019) developed a new uncertainty measure based on the predicted loss from the intermediate representations of the model. Although widely popular, purely uncertainty-based strategies relying on batch selection are susceptible to query samples with redundant information. However, manually annotating similar samples is a waste of annotation resources. Moreover, incorporating a set of similar samples to the labelled training set could bias the model towards an area outside the true data distribution. These samples could hence hamper rather than improve model generalization.

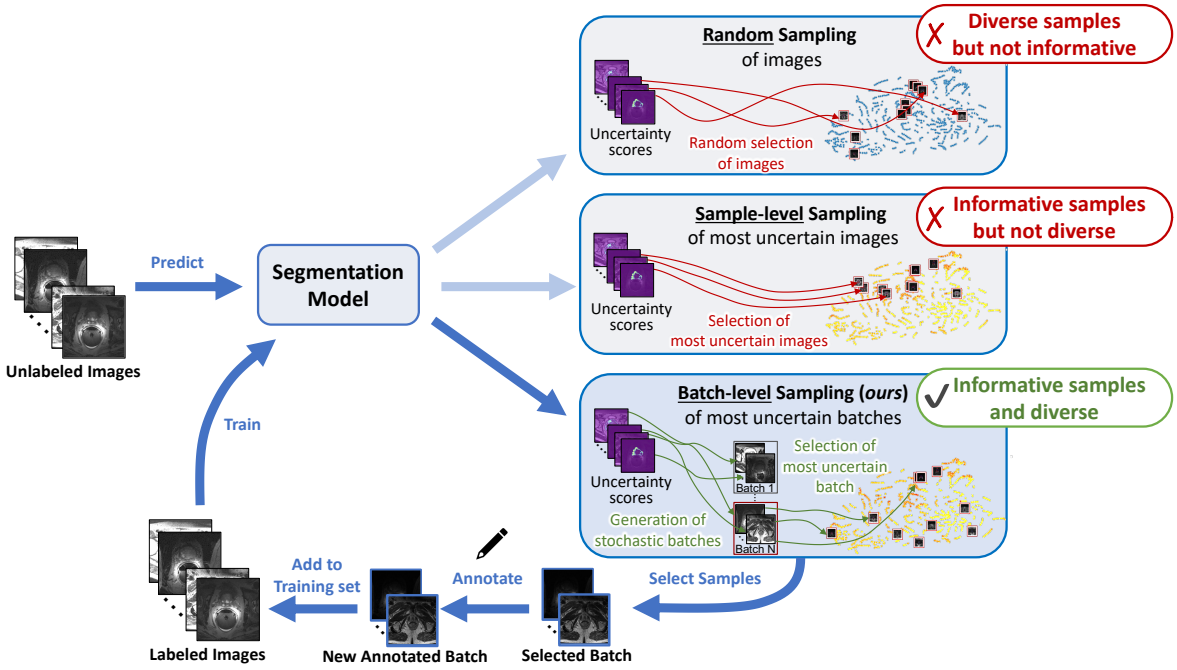


Fig. 1: **Stochastic batch AL for uncertainty-based sampling**. Our sampling method combines the diversity of random sampling with the informativeness of uncertainty-based sampling. Adding our stochastic batch paradigm enables the data uncertainty to be estimated in a broader *batch-level* selection rather than a *sample-level* selection. After selecting a candidate set of unlabelled samples, the set is annotated and added to the existing labelled set. Finally, the segmentation model is retrained.

2.2. *Representative-based AL methods*

As opposed to uncertainty-based approaches, representative-based AL methods aim at diversifying the batch of candidate samples to improve the future performance of the model (Settles, 2009). One of the main representative-based approaches, Core-set (Sener and Savarese, 2018), identifies the most diverse and representative samples by minimizing the distance between the latent representations of labelled and unlabelled images, as given by the task model. Core-set aims for the model to perform as well with the candidate set as it would with the entire dataset. While specifically designed to be applied to complex models such as Convolutional Neural Networks (CNNs), core-set selection does not scale well to high-dimensional data since it requires computing the Euclidean distance between all pairs of data samples. A later work, VAAL (Sinha et al., 2019), learns a smooth latent-state representation of the input data via a variational auto-encoder (VAE). VAAL then selects samples different from the ones already labelled based on the learnt latent representation. Since the VAE is task-agnostic, VAAL can, however, easily query outlier data. In addition, it provides no mechanism to avoid choosing overlapping samples and requires careful tuning of its added modules.

2.3. *Hybrid AL strategies*

Against the limitations of uncertainty-based methods, hybrid strategies try to find a balance between uncertainty and diversity measures to identify the most informative samples (Settles, 2009). They usually combine existing approaches. An early study proposed to adaptively choose the best AL strategies from a candidate set of methods (Hsu and Lin, 2015). However, most hybrid methods first compute model uncertainty before ensuring sample diversity through a similarity metric. For instance, Suggestive Annotation (Yang et al., 2017) applies core-set selection on a subgroup of the most uncertain samples obtained through bootstrapping. BADGE (Ash et al., 2020) uses gradient embeddings to account for uncertainty (uncertain samples will have a gradient embedding with higher norm) and employs Kmeans++ initialization on top of these embeddings to ensure the diversity of selected samples. Nath et al. (2021) combine prevailing mutual information and entropy measures to ensure diversity and optimize training by duplicating difficult samples. Observing that uncertainty-based approaches fail to exploit the data distribution and representative-based approaches are task-agnostic, Task-aware VAAL (Kim et al., 2021) incorporates the uncertainty measure proposed by the method Learning Loss (Yoo and Kweon, 2019) to VAAL's (Sinha et al., 2019) latent representation. While these studies rely on a two-step approach, Sourati et al. (2019) directly solve an optimization problem for batch-mode sampling, yielding a distribution of candidate samples rather than specific examples. However, just like representative-based AL strategies, most of these works are difficult to scale due to their computational complexity (Ash et al., 2020; Nath et al., 2021; Sourati et al., 2019; Yang et al., 2017). Alternatively, they may require external modules, which increase the range of parameters to tune and learn (Kim et al., 2021).

2.4. *AL for medical image segmentation*

High-dimensional data remains a particularly challenging problem in AL (Ren et al., 2021). Therefore, most studies

on AL applied to computer vision primarily focus on low-dimensional annotation tasks such as image classification (Gal et al., 2017; Wang et al., 2017; Sener and Savarese, 2018; Beluch et al., 2018; Sourati et al., 2019; Gao et al., 2020; Ash et al., 2020; Zhang et al., 2022). Moreover, approaches tackling pixel-wise annotations predominantly address the segmentation of natural images (Sinha et al., 2019; Huang et al., 2021; Kim et al., 2021; Xie et al., 2022).

Earlier work applying AL to medical image segmentation has relied on geometric priors to query planes or supervoxels of maximum uncertainty, without adopting deep learning-based models (Top et al., 2011; Konyushkova et al., 2015, 2019). One of the initial deep AL frameworks for this task, Suggestive Annotation (Yang et al., 2017), uses bootstrapping to estimate sample uncertainty and a greedy cosine similarity measure to evaluate the similarity between the candidate set and the unlabelled pool. Similarly, Li and Yin (2020) propose to select a candidate set with a high disagreement among the predictions of K models and a minimal discrepancy between the labelled and unlabelled sets. Instead of relying on multiple models, Ozdemir et al. (2018) employ a Bayesian network with Monte Carlo dropout to compute prediction variance, and adopt a Borda-count-based sampling strategy to find the best-ranked candidates in terms of uncertainty and representativeness. An extension of this approach instead computes the representativeness with an infoVAE (Zhao et al., 2019) for a maximum-likelihood sampling in the latent space (Ozdemir et al., 2021). Nath et al. (2021) build a mutual information-based metric, computed between the labelled and unlabelled pools, to ensure the diversity of the candidate set. However, these approaches tend to be computationally expensive and challenging to scale to large datasets. Instead of relying on a 2-step approach, Sourati et al. (2018) propose a method based on the Fisher information to directly solve an optimization problem that outputs a distribution to sample from. Alternative approaches have opted for membership query synthesis as an AL strategy, producing synthetic samples for annotation. For instance, Mahapatra et al. (2018) employ a conditional generative adversarial network (cGAN) to generate realistic-looking chest X-ray images conditioned on real images, and a Bayesian neural network to select which ones would be most informative when used as training data. Other approaches propose a sample selection strategy which also covers the initial labelled set (Smailagic et al., 2018; Nath et al., 2022; Li et al., 2023). Recently, a comparative study of existing strategies for 3D medical image segmentation found that random sampling and strided sampling served as particularly strong baselines for this type of task (Burmeister et al., 2022). The study also observed that representative-based strategies did not perform well in early stages, which the authors attribute to poor feature vectors generated by the model trained on very few labelled samples.

3. Methods

Given a labelled set $\mathcal{D}_L = \{(\mathbf{x}^{(j)}, \mathbf{y}^{(j)})\}_{j=1}^{N}$, with data $\mathbf{x} \in \mathbb{R}^{H\times W}$ and segmentation mask $\mathbf{y} \in \mathbb{R}^{C\times H\times W}$ (H and W are respectively the image height and width, and C is the number of classes), we train a fully-supervised segmentation model $f_\theta(\cdot)$ parameterized by θ with labelled samples from $\mathcal{D}_L$.

After training the model f_θ with $\mathcal{D}_L$ (corresponding to one training cycle), we select B samples from the unlabelled set $\mathcal{D}_U = \{\mathbf{x}_u^{(j)}\}_{j=1}^{M}$. These samples are annotated by an oracle before being added to the labelled training set $\mathcal{D}_L$. The new labelled and unlabelled sets are updated such that $|\mathcal{D}_L| = N + B$ and $|\mathcal{D}_U| = M - B$. This iterative process is repeated until the total annotation budget is exhausted.

Our AL method addresses the problem of uncertainty-based strategies, generally prone to query samples with redundant information, in a simple and computationally-efficient way. It builds upon our use of stochastic batches and operates in two stages to ensure a guided sampling diversity, summarized in Fig. 1. First, we generate a pool of Q batches, each containing B samples chosen uniformly at random from $\mathcal{D}_U$:

$$Batch^{(i)} = \{\mathbf{x}_u^{(i_1)}, \mathbf{x}_u^{(i_2)}, ..., \mathbf{x}_u^{(i_B)}\} \sim Uniform(\mathcal{D}_u, B) \tag{1}$$

For each generated batch, an uncertainty score is assigned to each unlabelled sample it contains, according to the current model $f_{\hat{\theta}}$ and the chosen uncertainty metric ($Uncert$):

$$\forall k = 1, ..., B: \quad u_{score}^{x_u^{(i_k)}} = Uncert\,(f_{\hat{\theta}}, \mathbf{x}_u^{(i_k)}). \tag{2}$$

The mean u_{score} is computed across each generated batch:

$$u_{score}^{Batch^{(i)}} = \frac{1}{B}\sum_{k=1}^{B} u_{score}^{x_u^{(i_k)}}. \tag{3}$$

The batch with the highest mean score yields the set of annotation candidates $X_{candidate}$, such that:

$$X_{candidate} \leftarrow \underset{Batch^{(i)}}{\operatorname{argmax}} \left(u_{score}^{Batch^{(i)}}\right). \tag{4}$$

The algorithm for our stochastic batch selection strategy is presented in Alg. 1.

Algorithm 1 Uncertainty-based sampling with Stochastic Batches

Input $\mathcal{D}_u$, Q, B
for $\mathbf{x}_u \in \mathcal{D}_u$ **do**
  $u_{score} \leftarrow Uncert(f_{\hat{\theta}}, \mathbf{x}_u)$
end for
for $i \leftarrow 1$ to Q **do**
  $Batch^{(i)} = \{\mathbf{x}_u^{(1)}, ..., \mathbf{x}_u^{(B)}\} \leftarrow Uniform(\mathcal{D}_u, B)$
  $u_{score}^{Batch^{(i)}} \leftarrow$ Mean u_{score} over all samples in $Batch^{(i)}$
end for
$X_{candidate} \leftarrow \operatorname{argmax}_{Batch^{(i)}} \left(u_{score}^{Batch^{(i)}}\right)$
Return $X_{candidate}$

4. Experiments

We assess the benefits of our proposed stochastic batches on a medical image segmentation task. Our evaluation compares the performance with and without stochastic batches of models

trained with different uncertainty-based AL strategies. These strategies include Entropy-based sampling (Shannon, 1948), Dropout-based sampling (Gal and Ghahramani, 2016), Test-time augmentation (TTA)-based sampling (Gaillochet et al., 2022) and sampling based on Learning Loss (Yoo and Kweon, 2019), defined in more details in Sec. 4.3.2. We start by evaluating the gains of our stochastic batch sampling on two medical image datasets. We then assess the robustness of our method to the training and sampling procedure through a series of ablation studies on the initial labelled set size, training hyper-parameters, sampling budget and stochastic pool size.

4.1. Datasets

We validate our method on two complementary datasets with different types of challenges: 1) the Prostate MR Image Segmentation (PROMISE) 2012 challenge (Litjens et al., 2014), for prostate segmentation, with varying degrees of pixel intensity distributions (as pictured in Fig. 6), and 2) the Medical Segmentation Decathlon (Antonelli et al., 2022) for the segmentation of anterior and posterior hippocampus, with varying degrees of anatomical shapes.

The PROMISE12 dataset contains MRI data from 50 patients, both healthy (or with benign diseases) and pathological (with prostate cancer). Each volume is converted to 2D images by slicing along the short axis. Images are then resampled to 1.0 mm isotropic resolution and resized to 128×128 pixels.

Similarly, the Medical Segmentation Decathlon contains hippocampus data from 260 patients. The MRI volumes are converted to 2D images, which are resized to 50×50 pixels while kept to the original 1.0 mm isotropic resolution. The pixel intensity of both datasets is normalized based on the 1% and 99% percentiles for each scan.

We test our model on 10 patient volumes from the prostate dataset and 50 from the hippocampus dataset, all selected uniformly at random. This yields 248 and 1757 test images, respectively. Our validation uses 109 prostate images composing 5 volumes, and 350 hippocampus images composing 10 volumes. Since active learning aims to minimize the amount of labelled data, we only use this validation set for hyper-parameter search purposes. Our ablation studies show that our method remains advantageous under different hyper-parameter settings. Our training set, labelled and unlabelled, comprises 1020 prostate images from 35 patients and 7163 hippocampus images from 200 patients.

4.2. Evaluation metrics

We evaluate our method on test volumes (3D) and individual images from these volumes (2D). We use both pixel overlap-based metrics and distance-based metrics.

In terms of overlap-based metrics, we use the well-known Dice similarity coefficient (DSC), which ranges from 0% (zero overlap) to 100% (perfect overlap):

$$\mathrm{DSC}(X, Y) = \frac{2|X \cap Y|}{|X| + |Y|} \tag{5}$$

In our results, we report the DSC averaged over all non-background channels.

The Hausdorff distance (HD) measures the quality of the segmentation by computing the maximum shortest distance between a point from the prediction contour and a point from the target contour. Since the Hausdorff distance tends to be sensitive to outliers, we use a more robust variant which considers the 95^{th} percentile instead of the true maximum (HD95). Given $\mathrm{d}(x, Y)$ the minimum distance from the boundary pixel x to the region Y, we get:

$$\mathrm{HD95}(X, Y) = \max\left\{95^{th}_{x \in X}\, \mathrm{d}(x, Y),\ 95^{th}_{y \in Y}\, \mathrm{d}(X, y)\right\} \tag{6}$$

4.3. Implementation details

Medical annotations for image segmentation are typically performed on all slices of a given image volume (Ozdemir et al., 2021). However, to optimize the limited annotation resources, we conduct slice-based active learning and select individual images for annotation after every cycle. We start each experiment by training our model with 10 labelled images, randomly sampled from the unlabelled set before annotation. Setting the budget to $B = 10$, we use our AL strategy to select 10 new samples from the unlabelled set, annotate them and add them to the existing labelled set. This process corresponds to the first AL cycle, which we repeat for a fixed number of cycles. Similarly to the experimental setting of previous studies, we retrain the model from scratch after each AL cycle to evaluate model performance in a consistent way (Budd et al., 2021).

Random processes such as model initialization or data shuffling are seeded. We repeat each experimental setup with 5 different seeds and report the mean and standard deviation of these runs as our result. Experiments were run on NVIDIA V100 and A6000 GPUs, with CUDA 10.2 and CUDA 12.0, respectively. We implement the methods using Python 3.8.10 with the PyTorch framework.

4.3.1. Training

State-of-the-art methods in medical image segmentation have often adopted UNet-based architectures (Ronneberger et al., 2015). Accordingly, we use a standard 4-layer UNet as a proxy for widely used architectures in our segmentation model, with dropout ($p = 0.5$), batch normalization and a leaky ReLU activation function. Employing such a model also focuses the evaluation on the improvement due to our stochastic batch strategy instead of measuring the performance of a backbone. However, without loss of generality, the use of alternative segmentation models could also be envisioned for our AL approach.

The model is trained for 75 epochs in all experiments, each iterating over 250 batches (training samples can appear in several batches), with a batch size of 4. Training is hence carried out for a fixed $75 \times 250 = 18,750$ steps in all experiments, ensuring a fairer comparison of model performance between AL cycles.

We optimize a supervised CE loss with the Adam optimizer (Kingma and Ba, 2015). We apply a gradual warmup with a cosine annealing scheduler (Loshchilov and Hutter, 2017; Goyal et al., 2018) to control the learning rate. During training, we use data augmentations on the input, with parameters d and ϵ,

where d is the degree of rotation in 2D, and ϵ models Gaussian noise.

When not testing for their impact, we keep the training hyperparameters fixed. We fix the initial learning rate $LR = 10^{-6}$ with optimizer weight decay set to 10^{-4}. The scheduler increments the learning rate by a factor 200 during the first 10 epochs. For augmentations, we set $d \sim \mathcal{U}(-10, 10)$ and $\epsilon \sim \mathcal{N}(0, 0.01)$.

Since active learning aims to minimize the amount of labelled data needed to train the model, we minimize the use of the validation set and avoid its use to select the final model. Our final model is instead the model obtained after the last training epoch.

4.3.2. Active learning sampling

Baselines. We compare our stochastic batches strategy with random sampling (RS), Core-set (Sener and Savarese, 2018), and four purely uncertainty-based methods:

- Entropy-based uncertainty (Shannon, 1948), which computes the entropy on the predicted output probabilities:

$$Uncert\,(f_{\hat{\theta}}, \mathbf{x}_u^{(i_k)}) = -\sum_i p(y_i|\mathbf{x}_u^{(i_k)}, \hat{\theta})\ \log p(y_i|\mathbf{x}_u^{(i_k)}, \hat{\theta});$$

- Dropout-based uncertainty (Gal and Ghahramani, 2016), using the divergence of K predictions obtained by multiple inferences with dropout d:

$$Uncert\,(f_{\hat{\theta}}, \mathbf{x}_u^{(i_k)}) = Div(f_{\hat{\theta},d_1}(\mathbf{x}_u^{(i_k)}),\ ...,\ f_{\hat{\theta},d_K}(\mathbf{x}_u^{(i_k)}));$$

- Test-time Augmentation (TTA)-based uncertainty (Gaillochet et al., 2022), which measures the divergence of predictions obtained for K transformations Γ to the input:

$$Uncert\,(f_{\hat{\theta}}, \mathbf{x}_u^{(i_k)}) = Div\,(\Gamma_1^{-1}[f_{\hat{\theta}}(\Gamma_1(\mathbf{x}_u^{(i_k)}))],\ ...,\ \Gamma_K^{-1}[f_{\hat{\theta}}(\Gamma_K(\mathbf{x}_u^{(i_k)}))]);$$

- Learning Loss uncertainty (Yoo and Kweon, 2019), which trains an external module $L_{\tilde{\theta}}$ to predict the target losses from a feature set h extracted from the hidden layers of $f_{\hat{\theta}}$:

$$Uncert\,(f_{\hat{\theta}}, L_{\tilde{\theta}}, \mathbf{x}_u^{(i_k)}) = L_{\tilde{\theta}}(h(\mathbf{x}_u^{(i_k)})).$$

These purely uncertainty-based methods query batches made of the most uncertain samples according to a sample-level uncertainty measure.

Similarly to Gaillochet et al. (2022), as our divergence measure for Dropout-based and TTA-based sampling, we use a standard Jensen–Shannon divergence (JSD) on the output probability maps obtained from $K = 8$ inferences. For TTA, augmentations Γ include Gaussian noise $\epsilon \sim \mathcal{N}(0, 0.01)$ and rotation. To simulate more realistic transformations in medical data, we replace the 90 degrees rotations in Gaillochet et al. (2022) with rotations of angle $d \sim \mathcal{U}(-10, 10)$ degrees. The training parameters used for the approach based on Learning Loss (Yoo and Kweon, 2019) were obtained by grid search on 10 labelled samples. We kept these parameters fixed in all our experiments.

Stochastic batches. We generate the pool of stochastic batches by iteratively sampling B unlabelled images uniformly at random and without replacement. In other words, we divide the unlabelled samples into Q pools of B samples. Hence the stochastic pool has size $Q = \text{floor}(|\mathcal{D}_U|/B)$, and it reduces in size with the number of AL cycles.

5. Results

5.1. AL performance on the Prostate and Hippocampus datasets

We validate our proposed stochastic batch sampling strategy by looking at the AL performance over 5 different initial labelled sets chosen uniformly at random from the training set. Tab. 1 shows the average results over all AL cycles for both Prostate and Hippocampus data. Note that the standard deviations given in the table tend to be large because they are averaged over multiple initial labelled sets, initialization seeds and AL cycles. For all methods and metrics except for TTA on Prostate with the 95% Hausdorff distance metric, stochastic batch sampling constantly provides improved performance over its purely uncertainty-based counterpart, both in terms of overlap-based and distance-based metric.

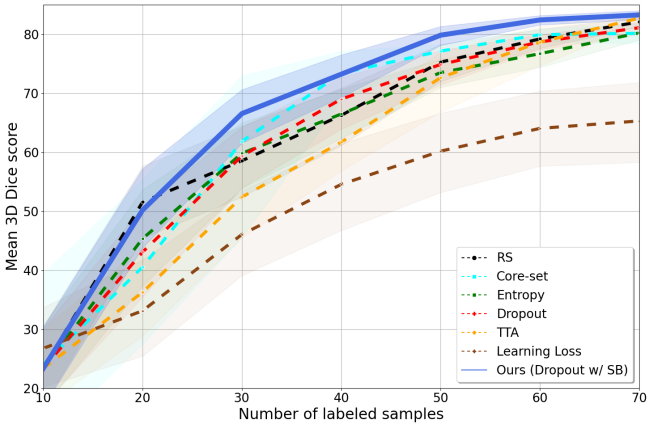


Fig. 2: **Overall AL performance on the Prostate dataset**. Our best stochastic batch sampling method (full-blue) outperforms all other methods, including Core-set and random sampling (RS).

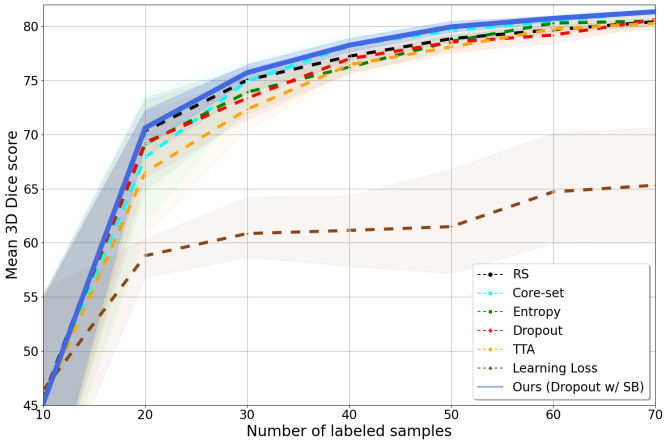


Fig. 3: **Overall AL performance on the Hippocampus dataset**. Our best stochastic batch sampling method (full-blue) outperforms all other methods, including Core-set and random sampling (RS).

We also observe that stochastic batch sampling outperforms both random sampling and Core-set (Sener and Savarese, 2018), a diversity-based AL approach. This is corroborated by

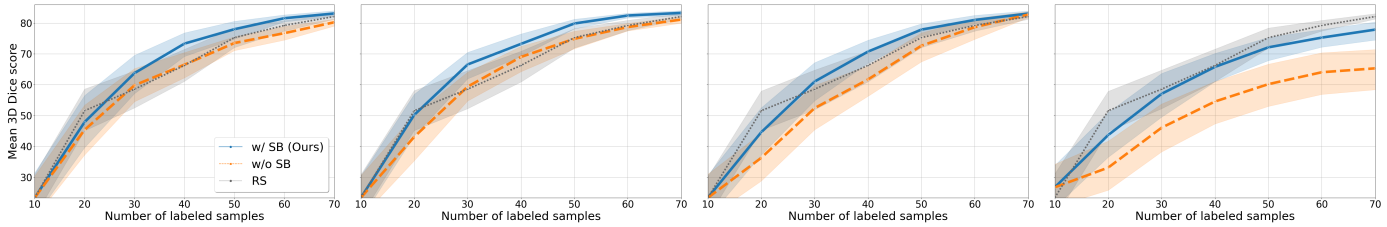
Table 1: **Overall improvements with Stochastic Batches over varying initial labelled samples**. Mean model performance over all AL cycles. We show the mean (std) Dice score (DSC, higher is better) and 95% Hausdorff distance (HD95, lower is better) over 3D test volumes and 2D test images. The results are averaged over 5 initial labelled sets chosen uniformly at random and 6 AL cycles (we omit results with the initial labelled set as they are similar across all methods). A * indicates the statistical significance of the result with a p-value < 0.05 given a paired permutation test.

		Prostate			**Anterior Hippocampus**			**Posterior Hippocampus**		
		3D DSC	**2D DSC**	**3D HD95**	**3D DSC**	**2D DSC**	**3D HD95**	**3D DSC**	**2D DSC**	**3D HD95**
RS		68.83 (±15.99)	67.94 (±8.28)	7.032 (±3.734)	77.42 (±1.67)	75.45 (±1.13)	4.09 (±0.47)	76.43 (±0.80)	**70.02** (±1.62)	4.51 (±0.70)
Core-set (Sener and Savarese, 2018)		68.84 (±17.37)	65.87 (±7.31)	7.64 (±2.73)	78.83 (±3.25)	73.14 (±1.20)	4.45 (±0.46)	75.32 (±5.46)	66.45 (±1.29)	4.52 (±0.53)
Entropy (Shannon, 1948)	w/o SB	67.01 (±16.68)	66.88 (±8.62)	7.026 (±4.271)	78.22 (±1.90)	75.03 (±0.97)	3.79 (±0.23)	74.68 (±1.60)	65.70 (±1.66)	5.10 (±1.10)
	Ours	71.27* (±17.39)	68.99* (±9.03)	6.689* (±3.143)	79.25* (±0.86)	75.84 (±0.86)	**3.72** (±0.15)	76.23* (±0.87)	69.01* (±1.97)	**3.85*** (±0.31)
Dropout (Gal and Ghahramani, 2016)	w/o SB	67.69 (±17.16)	67.07 (±9.51)	6.964 (±4.952)	78.22 (±1.28)	74.29 (±1.10)	4.04 (±0.33)	74.45 (±1.20)	66.78 (±2.06)	4.77 (±1.19)
	Ours	**72.59*** (±14.96)	**69.64*** (±8.05)	**6.583*** (±3.177)	**79.28*** (±0.83)	**76.36*** (±0.69)	3.73 (±0.10)	76.27* (±0.85)	**68.94*** (±1.15)	**3.88*** (±0.39)
TTA (Gaillochet et al., 2022)	w/o SB	64.07 (±21.13)	65.85 (±10.25)	6.918* (±4.794)	77.31 (±3.24)	73.66 (±1.08)	4.10 (±0.54)	73.84 (±2.01)	64.94 (±0.59)	5.07 (±0.93)
	Ours	69.71* (±17.59)	68.00* (±9.02)	7.188 (±3.173)	78.86* (±0.94)	75.25 (±1.41)	4.07 (±0.31)	**76.44*** (±0.90)	67.08* (±1.01)	4.43* (±0.40)
Learning Loss (Yoo and Kweon, 2019)	w/o SB	53.88 (±21.51)	60.22 (±10.36)	9.139 (±6.439)	62.54 (±1.38)	69.70 (±0.92)	5.94 (±0.59)	61.57 (±2.83)	62.82 (±1.14)	5.87 (±0.12)
	Ours	65.29* (±17.72)	65.72* (±8.94)	7.816* (±4.384)	72.09* (±2.73)	74.32* (±0.85)	4.51* (±0.60)	71.23* (±1.29)	67.75* (±1.21)	5.16 (±0.63)

Fig. 2 and Fig. 3, which show that Dropout with our stochastic batches outperforms all other baseline methods in terms of 3D dice score, in almost all AL cycles. In addition, Tab. 2 gives the average time required by each strategy to provide a candidate set for annotation from the Hippocampus dataset. We see that using stochastic batches does not increase the sampling time of uncertainty-based methods. Furthermore, sampling with our proposed method is always much faster than with Core-set.

When looking at each dataset in more detail, the pairwise results on the Prostate dataset, shown in Fig. 4, validate the effectiveness of our method against different initial labelled sets. Averaged over 25 experiments with varying initial labelled sets and initialization seeds, our stochastic batch querying (blue, full lines) improves the model's performance of purely uncertainty-based strategies (orange, dashed lines). For all considered AL strategies, selecting the most uncertain batch of samples rather than the most uncertain individual samples improves the model's overall performance. The 3D dice score is always boosted, either over the score obtained by random sampling (grey, dotted) or to a level similar to that of a random sampling if the score were originally much lower, such as in the case of the Learning Loss. Indeed, Learning Loss has noticeably lower performance compared to Entropy, Dropout and TTA-based sampling. The Learning Loss approach involves backpropagating the gradient through both the task model and loss module during training. Both are updated simultaneously, which means that training the loss module affects the training of the task model and vice versa. For comparability reasons and following most works in AL, we tuned the hyper-parameters such that the best validation performance was obtained on the first AL cycle (with the initial labelled set). We believe this could explain the poorer performance of Learning Loss with an increasing number of labelled samples. Similar observations can be made from Hippocampus data, as shown in Fig. 5.

We also visually investigate the benefits of using our stochastic batches with an uncertainty-based sample selection. In Fig. 6, we show two sets of candidate samples from the Prostate dataset identified by Entropy-based sampling, with and without our stochastic batches. The first two columns show samples selected by identifying the most uncertain randomly generated batch. The last two columns depict the most certain queried samples based on the individual entropy of their predicted output probabilities. While the samples from the first two columns seem more diverse, with more variety in the candidate set, the third column contains nearly identical samples. Indeed, tracking the first four images of the column to their corresponding 3D volume shows that the slices were taken from the MRI volume of the same patient. This confirms our claim that purely uncertainty-based strategies are likely to select very similar samples and that our stochastic batch sampling reduces



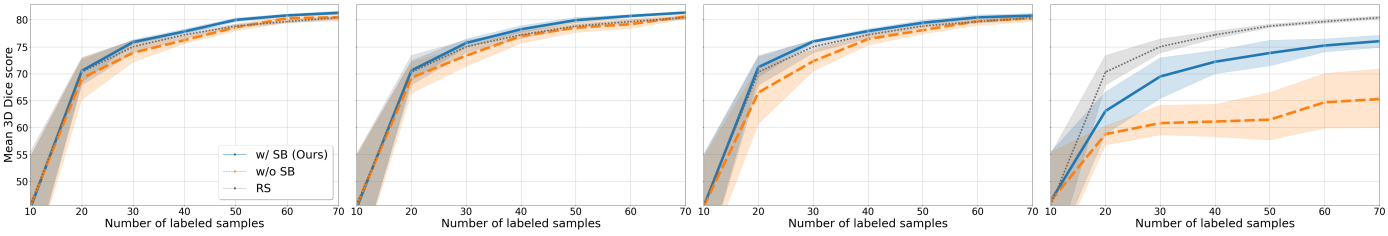
(a) Improvements for Entropy (Shannon, 1948)

(b) Improvements for Dropout (Gal and Ghahramani, 2016)

(c) Improvements for TTA (Gaillochet et al., 2022)

(d) Improvements for Learning Loss (Yoo and Kweon, 2019)

Fig. 4: **Individual improvements with Stochastic Batches on the Prostate dataset**. Active learning results in terms of 3D test dice score and corresponding 95% confidence interval. The results are averaged over 5 different initial labelled sets and 5 initialization seeds. Depicted are the results for sampling based on a) Entropy, b) Dropout, c) Test-time augmentation and d) Learning Loss). The active learning selection is shown with (blue, full) and without (orange, dashed) stochastic batches, and random sampling is plotted in dotted grey. Stochastic batches improve the model performance of purely uncertainty-based AL strategies, regardless of the initial labelled set, repeatedly outperforming random sampling.



(a) Improvements for Entropy (Shannon, 1948)

(b) Improvements for Dropout (Gal and Ghahramani, 2016)

(c) Improvements for TTA (Gaillochet et al., 2022)

(d) Improvements for Learning Loss (Yoo and Kweon, 2019)

Fig. 5: **Individual improvements with Stochastic Batches on the Hippocampus dataset**. Active learning results on the Hippocampus dataset in terms of 3D test dice score and corresponding 95% confidence interval. The results are averaged over 5 different initial labelled sets. Depicted are the results for sampling based on a) Entropy, b) Dropout, c) Test-time augmentation and d) Learning Loss. Sampling with Stochastic batches (blue, full) improves the model performance of purely uncertainty-based AL strategies (orange, dashed), regardless of the initial labelled set, boosting it above random sampling (grey, dotted) in the majority of cases.

the probability of querying samples with highly overlapping information.

Finally, we examine the impact of our selection strategy on the segmentation of test data. In Fig. 7, we see that the model trained on images selected via our stochastic batch sampling method outputs better anterior and posterior hippocampus segmentations. By the fourth cycle, the segmentation reaches a mean DSC (over both classes) of 81.15%, compared to the 68.03% obtained via a purely Entropy-based sampling.

5.2. *Ablation experiments on the Prostate dataset*

To evaluate the robustness of our method to different experimental settings, we perform a series of ablation studies on the Prostate dataset, evaluating the impact of the initial labelled set size, training hyper-parameters, sampling budget and stochastic pool size.

5.2.1. *Impact of initial labelled set size*

For our first ablation study, we validate the performance of models trained on initial labelled sets of varying sizes. For each given initial labelled set size, the experiment is repeated with 5 initialization seeds controlling the initial labelled samples used, the model initialization and the training updates. Table 3 gives the average model performance over 6 AL cycles. We observe that our stochastic batch selection strategy improves upon purely uncertainty-based selection also when we vary the initial number of labelled samples.

5.2.2. *Impact of training hyper-parameters*

Active Learning methods typically tune hyper-parameters using an initial labelled set, maintaining these settings throughout all AL cycles. However, these parameters might be suboptimal for subsequent training cycles as more labelled data becomes available. We hence explore the robustness of stochastic batches to different yet realistic training hyperparameters. We

Table 2: **Sampling time.** Mean sampling time computed over all AL cycles, for the Hippocampus dataset.

	RS	Core-set	Entropy		Dropout		TTA		Learning Loss	
			w/o SB	Ours	w/o SB	Ours	w/o SB	Ours	w/o SB	Ours
Time (min.)	0.00	0.71	0.12	0.11	0.58	0.58	0.37	0.37	0.16	0.18

select five hyperparameter sets, each optimized for labelled set sizes of 10, 50, 100, 150, and 200. These sets included diverse augmentation parameters, scheduling parameters and loss function weights.

Results in Fig. 8 reveal that our stochastic batch sampling noticeably improves the performance of purely uncertainty-based sampling, particularly in the first 3 or 4 AL cycles. In addition, the spread of 3D dice scores tends to be narrower with our method than with a purely uncertainty-based sampling, showing that our strategy tends to be more stable.

The benefit of using our stochastic batches is most evident in the average dice scores over all AL cycles for both test images and volumes, as given in Tab. 4. Test-Time Augmentation (TTA) generally performs better with stochastic batches, although the results are not statistically significant for distance-based metrics. This could be due to the fact that we vary the training and regularization hyper-parameters while keeping data augmentation parameters fixed for sampling.

5.2.3. Impact of sampling budget

We also investigate the robustness of stochastic batches to the sampling budget. Keeping the initial labelled set and training hyper-parameters fixed, we run experiments with 5 different sampling budgets, which we keep constant across cycles. In this experiment, since we vary B, images are allowed to be resampled when generating the stochastic batches, and we keep the number of generated batches to a fixed $Q = 100$.

The results shown in Fig. 9 reveal that stochastic batches have a more consistent impact on model performance as the budget size increases. With a high budget $B = 15$, the use of stochastic batches constantly improves purely uncertainty-based methods. An improvement is also visible for lower budgets, such as $B = 5$, particularly for the Entropy, Dropout and TTA-based sampling.

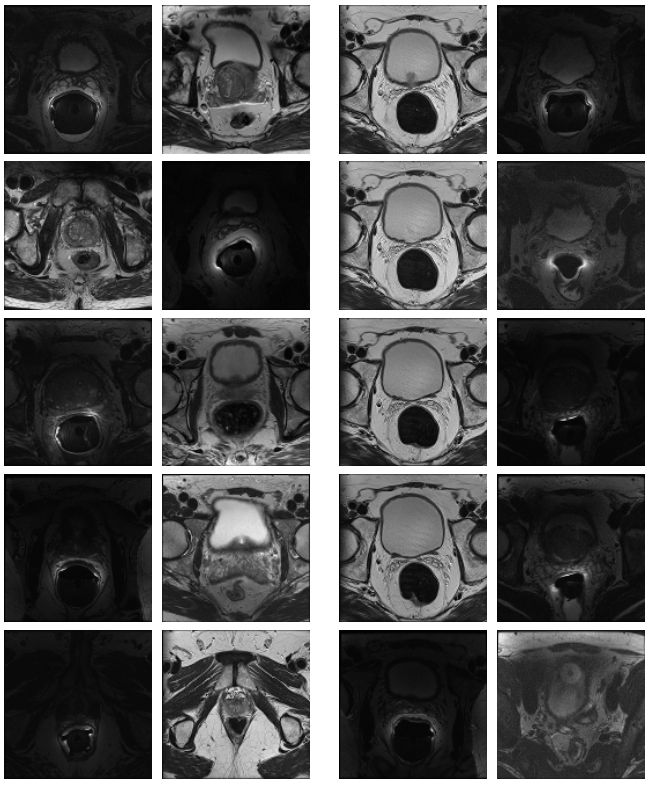
However, with very low budgets, batch uncertainty is highly influenced by the uncertainty of each individual sample, potentially reducing the benefits of diversity offered by stochastic

Table 3: **Overall improvements with Stochastic Batches for initial labelled sets of different sizes**. Mean model performance on the Prostate data over all AL cycles for initial sets of different sizes. We show the mean (std) Dice score (higher is better) over 3D test volumes (3D DSC). The results are averaged over 6 AL cycles (we omit results for the first AL cycle since all strategies share the same initial set). A * indicates the statistical significance of the result with a p-value < 0.05 given a paired permutation test.

	RS	Entropy (Shannon, 1948)		Dropout (Gal and Ghahramani, 2016)		TTA (Gaillochet et al., 2022)		Learning Loss (Yoo and Kweon, 2019)	
		w/o SB	Ours	w/o SB	Ours	w/o SB	Ours	w/o SB	Ours
5 initial samples	71.22 (±15.09)	65.18 (±14.43)	**72.36*** (±16.19)	61.69 (±18.15)	**71.45*** (±15.41)	64.16 (±16.43)	**67.67** (±16.70)	55.06 (±21.90)	**66.85*** (±19.30)
10 initial samples	71.08 (±11.70)	66.21 (±13.79)	**73.78*** (±10.73)	65.09 (±15.63)	**73.30*** (±14.95)	70.23 (±12.35)	**73.01** (±12.24)	48.51 (±9.83)	**60.73*** (±9.10)
15 initial samples	75.21 (±7.27)	0.7319 (±7.54)	**74.21** (±7.85)	72.90 (±6.96)	**76.81*** (±8.34)	**72.00** (±10.86)	71.53 (±8.84)	58.19 (±12.09)	**72.84*** (±10.32)
20 initial samples	76.00 (±7.19)	76.13 (±5.55)	**80.24*** (±4.19)	77.47 (±6.38)	**79.81*** (±05.54)	74.59 (±10.15)	**78.04** (±7.89)	69.81 (±6.74)	**75.51*** (±6.14)
25 initial samples	77.07 (±4.39)	77.73 (±3.79)	**79.71*** (±4.37)	77.44 (±4.31)	**81.08*** (±5.20)	76.81 (±9.03)	**78.32** (±5.88)	73.61 (±5.27)	**77.65*** (±5.52)

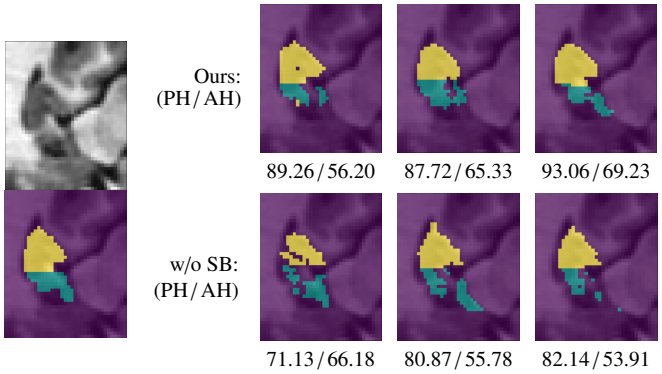
Table 4: **Overall improvements with Stochastic Batches over varying training hyper-parameters**. Mean model performance on Prostate data over all AL cycles (omitting training with the initial labelled set). We show the mean (std) Dice score (DSC, higher is better) and 95% Hausdorff (HD95, lower is better) distance over 3D test volumes and individual 2D test images. The results are averaged over 7 AL cycles and 5 training hyper-parameter sets. * indicates the statistical significance of the result with a p-value < 0.05 given a paired permutation test.

	RS	Entropy (Shannon, 1948)		Dropout (Gal and Ghahramani, 2016)		TTA (Gaillochet et al., 2022)		Learning Loss (Yoo and Kweon, 2019)	
		w/o SB	Ours	w/o SB	Ours	w/o SB	Ours	w/o SB	Ours
3D DSC (↑ best)	75.57 (±6.48)	75.13 (±6.95)	**78.44*** (±6.02)	76.49 (±7.65)	**78.59*** (±6.09)	77.33 (±6.92)	**78.67*** (±5.53)	69.53 (±8.43)	**76.25*** (±6.68)
2D DSC (↑ best)	68.29 (±6.79)	68.90 (±7.34)	**71.04*** (±6.51)	69.62 (±6.70)	**71.08*** (±6.79)	70.46 (±7.05)	**71.31*** (±5.71)	64.27 (±7.23)	**69.16*** (±6.80)
3D HD95 (↓ best)	7.58 (±3.86)	7.87 (±4.28)	**6.83*** (±3.31)	**6.72** (±2.75)	6.74 (±3.29)	6.32 (±2.87)	**6.13** (±2.82)	8.78 (±4.22)	**7.85*** (±3.68)



(a) Ours (b) w/o Stochastic Batches

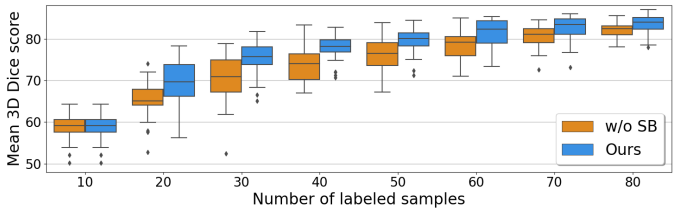
Fig. 6: **Candidate batches from the Prostate dataset**. The samples were selected by Entropy-based AL sampling with (first two columns) and without (last two columns) stochastic batches. While the candidate batch obtained via purely uncertainty-based sampling contains similar samples, selection with stochastic batches reduces the number of redundancies.



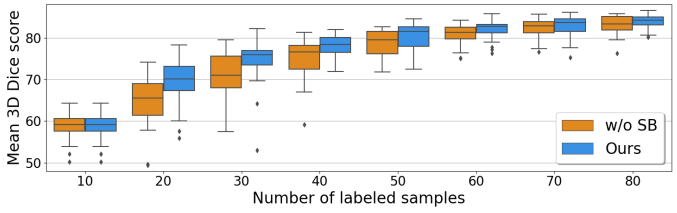
(a) Data & Target (b) 2^{nd} cycle (c) 3^{rd} cycle (d) 4^{th} cycle

Fig. 7: **Segmentation of a Hippocampus test sample across AL cycles**. The 2D dice score (DSC) is given for each predicted segmentation, both for the posterior hippocampus (PH, yellow) and the anterior hippocampus (AH, blue). At every AL cycle, the model trained on labelled samples selected with our stochastic batches (top row) predicts segmentations closer to the target mask (leftmost) compared to its purely Entropy-based counterpart (bottom row).

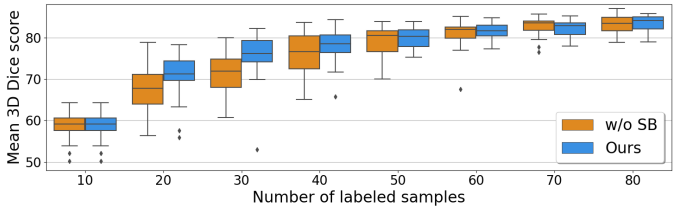
batches. The selection is dominated by uncertainty, and if the measure for uncertainty is not representative of the true uncertainty of the model, then uninformative samples could be selected and consequently bias the model.



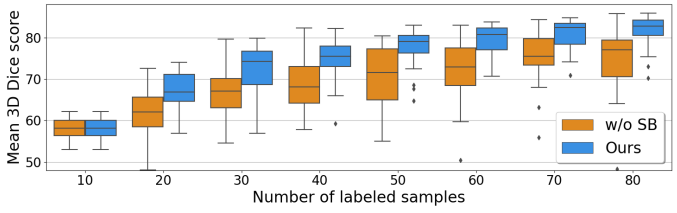
(a) Improvements for Entropy (Shannon, 1948)



(b) Improvements for Dropout (Gal and Ghahramani, 2016)



(c) Improvements for TTA (Gaillochet et al., 2022)

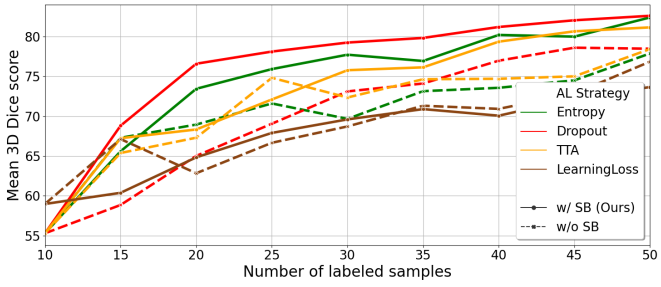


(d) Improvements for Learning loss (Yoo and Kweon, 2019)

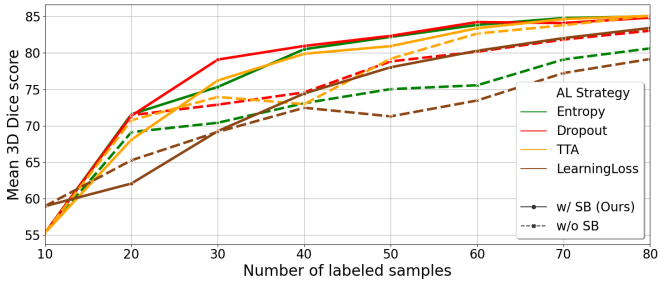
Fig. 8: **Improvements with Stochastic Batches over varying hyper-parameters**. Box plot of active learning results on Prostate data in terms of 3D test dice score, given over 5 training hyper-parameters sets and 5 initialization seeds. Depicted are the results for sampling based on a) Entropy, b) Dropout, c) Test-time augmentation and d) Learning loss. The AL selection is shown with (blue) and without (orange) stochastic batches. Our stochastic batches improve the model performance of purely uncertainty-based AL strategies and boost performance, even with variations in hyper-parameters.

5.2.4. Impact of sampling stochastic pool size

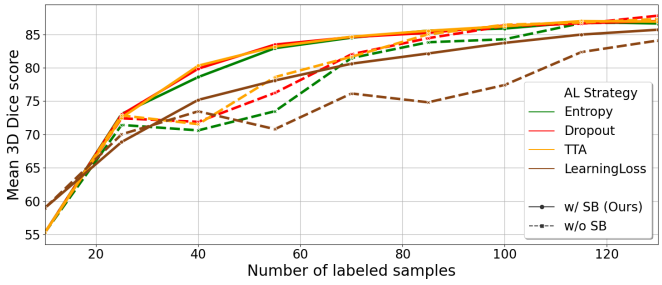
In our last ablation study, we evaluate the influence of the number of batches in the stochastic pool on the model performance, fixing the initial labelled set, training hyper-parameters and sampling budget. Instead of generating $Q = \text{floor}(|\mathcal{D}_u|/B)$ batches, we artificially vary Q. Accordingly, we allow resampling so samples can appear in multiple generated batches. The results for our experiments on Entropy-based and Dropout-based sampling are given in Fig. 10. Applying the biggest pool size does not necessarily yield the best performance. On the contrary, the model performs best when the most uncertain batch is selected from a pool containing 10 or 100 different batches. Increasing the pool of choices by 10 or 100 does not lead to significant improvements and can lead to worse perfor-



(a) Improvements with low budget ($B = 5$)

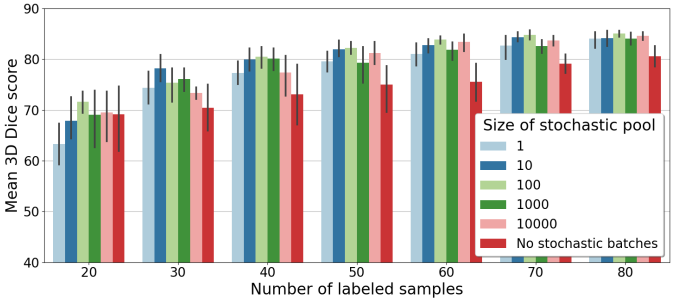


(b) Improvements with mid budget ($B = 10$)

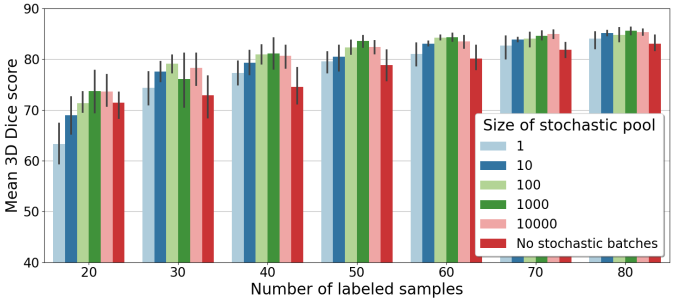


(c) Improvements with high budget ($B = 15$)

Fig. 9: **Improvements with Stochastic Batches given different budget sizes**. Model performance in terms of 3D dice score on test volumes given active learning selection with (solid) and without (dashed) stochastic batches on Prostate data. The results are given for sampling budgets a) $B = 5$, b) $B = 10$ and c) $B = 15$. Depicted are the results for sampling based on Entropy (green) and Dropout (red), TTA (yellow) and Learning Loss (brown). Using stochastic batches during sampling improves the model performance at both low and higher budgets.



(a) Stochastic batches for Entropy (Shannon, 1948)



(b) Stochastic batches for Dropout (Gal and Ghahramani, 2016)

Fig. 10: **Impact of pool size of Stochastic Batches**. Model performance in terms of 3D dice score on test volumes from Prostate data given stochastic batch pools of different sizes. The error bars (black) corresponds to the 95% confidence interval over 5 experiments with different seed initialization. Depicted are the results 2 popular uncertainty-based AL methods: Entropy-based sampling (10a) and Dropout-based sampling (10b). A medium pool size between 10 to 100 yields some of the most advantageous performances.

mances.

6. Discussion

Overall, our results demonstrate that using stochastic batches during uncertainty-based sampling is an efficient strategy to ensure diversity among the selected batch of samples. Furthermore, we experimentally observe that the benefit of using stochastic batches is robust to changes in the initial labelled set, initialization of the model and training hyper-parameters, as well as to variations in the sampling budget.

As illustrated in Fig. 6, the redundancy of queried samples constitutes one of the main drawbacks of uncertainty-based AL strategies. Their queried samples may indeed convey highly similar information. Hence, the annotation effort on these samples will be suboptimal. If, on the contrary, the most uncertain batches rather than the most uncertain samples are queried, the added diversity within our stochastic batches mitigates the overlap of information and redundancy between samples. Our stochastic scheme adds diversity to the uncertainty-based sampling in AL in a fast, computationally-efficient way, as shown by Tab. 2. Our quantitative results demonstrate the advantages of adding such a stochastic scheme in AL in terms of added segmentation accuracy in a low-labelled set regime and reduced number of required training samples.

Previous AL works have observed that the initial labelled pool can significantly impact the training and final performance of AL models (Chen et al., 2022). Nevertheless, a robust AL method should still perform well regardless of this initial labelled set. The results obtained in our experiment with varying initial labelled sets (Sec. 5.1 and Sec. 5.2.1) reveal that the performance boost from our stochastic batch sampling strategy is robust to changes in both the initial labelled set and model initialization. On average, selecting the most uncertain batches across AL cycles yields better results than selecting the most uncertain samples. Similarly, Sec. 5.2.2 shows that the improvements yielded by stochastic AL batches are also robust to changes in the training and regularization parameters. Hence, our method can maintain efficiency despite changes in the learn-

ing environment. These results suggest that using stochastic batches during AL for uncertainty-based sampling can be a reliable and robust AL approach.

Our stochastic batch querying strategy for uncertainty-based AL operates as a balance between a fully random and a purely uncertainty-based selection. While we set $Q = \text{floor}(|\mathcal{D}_U|/B)$, the stochastic pool size Q can also be directly modified to control the amount of randomness desired in the AL selection. With the smallest pool size ($Q = 1$), our stochastic batch selection is equivalent to random sampling since the single suggested batch will automatically have the highest uncertainty score in the pool. With the biggest pool size ($Q \to \infty$), all possible combinations of samples are available in the pool, and selecting the most uncertain batch of samples is equivalent to selecting the top uncertain samples. In other words, the approach becomes a purely uncertainty-based AL strategy with a larger pool size. As shown in Sec. 5.2.4, the benefits of our stochastic batches are apparent in between those extreme Q values, when the sampling strategy combines the informativeness of uncertainty-based sampling with the diversity provided by random sampling. Active learning is an expensive framework to experiment with, given that AL cycles are iterative and that procedures should be repeated to reduce as much as possible the influence of initialization. In this work, we ran multiple experiments with different settings (size and type of initial labelled set, training hyper-parameters, stochastic pool size, sampling budget) to test how stable our method was. However, we acknowledge that our experiments do not cover all ranges of possible setups.

7. Conclusion

Active learning is particularly relevant in medical image segmentation since manual labelling is highly time-consuming and expensive. This paper addresses three main limitations of AL strategies: the relatively limited literature on AL work for medical image segmentation compared to classification tasks, the tendency of uncertainty-based batch sampling strategies to select very similar samples and the computational burden of diversity-based methods. Instead of employing sample-level uncertainty for candidate selection, we suggest a batch-level approach where uncertainty is computed over randomly generated batches of samples. Using stochastic batches with uncertainty-based sampling is a simple, computational-inexpensive approach to improve the AL candidate selection and, hence, the final model performance. Our method is flexible and easily adaptable to any uncertainty-based AL strategy. In addition, our extensive experiments show that adding stochastic batches improves purely uncertainty-based methods consistently across different experimental setups. Hence, stochastic batching could bring a more reliable advantage over other representative-based works, which have shown significantly varying amounts of robustness in performance (Munjal et al., 2022). Our method could therefore act as a strong baseline to better use the limited annotation time of clinical experts when segmenting medical images.

Acknowledgments

This work is supported by the Canada Research Chair on Shape Analysis in Medical Imaging, the Research Council of Canada (NSERC) and the Quebec Bio-Imaging Network (QBIN). Computational resources were partially provided by Compute Canada. The authors also thank the PROMISE12 and the Medical Segmentation Decathlon challenge organizers for providing the data.

References

Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., Ronneberger, O., Summers, R.M., van Ginneken, B., Bilello, M., Bilic, P., Christ, P.F., Do, R.K.G., Gollub, M.J., Heckers, S.H., Huisman, H., Jarnagin, W.R., McHugo, M.K., Napel, S., Pernicka, J.S.G., Rhode, K., Tobon-Gomez, C., Vorontsov, E., Meakin, J.A., Ourselin, S., Wiesenfarth, M., Arbeláez, P., Bae, B., Chen, S., Daza, L., Feng, J., He, B., Isensee, F., Ji, Y., Jia, F., Kim, I., Maier-Hein, K., Merhof, D., Pai, A., Park, B., Perslev, M., Rezaiifar, R., Rippel, O., Sarasua, I., Shen, W., Son, J., Wachinger, C., Wang, L., Wang, Y., Xia, Y., Xu, D., Xu, Z., Zheng, Y., Simpson, A.L., Maier-Hein, L., Cardoso, M.J., 2022. The Medical Segmentation Decathlon. Nat Commun 13, 4128.

Ash, J.T., Zhang, C., Krishnamurthy, A., Langford, J., Agarwal, A., 2020. Deep Batch Active Learning by Diverse, Uncertain Gradient Lower Bounds, in: Eighth International Conference on Learning Representations (ICLR).

Beluch, W.H., Genewein, T., Nurnberger, A., Kohler, J.M., 2018. The Power of Ensembles for Active Learning in Image Classification, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Bengar, J.Z., van de Weijer, J., Twardowski, B., Raducanu, B., 2021. Reducing Label Effort: Self-Supervised meets Active Learning, in: IEEE/CVF International Conference on Computer Vision Workshops (ICCVW).

Budd, S., Robinson, E.C., Kainz, B., 2021. A survey on active learning and human-in-the-loop deep learning for medical image analysis. Medical Image Analysis 71, 102062.

Burmeister, J.M., Rosas, M.F., Hagemann, J., Kordt, J., Blum, J., Shabo, S., Bergner, B., Lippert, C., 2022. Less Is More: A Comparison of Active Learning Strategies for 3D Medical Image Segmentation, in: ICML Workshop on Adaptive Experimental Design and Active Learning in the Real World (ICML ReALML).

Chen, L., Bai, Y., Huang, S., Lu, Y., Wen, B., Yuille, A.L., Zhou, Z., 2022. Making Your First Choice: To Address Cold Start Problem in Vision Active Learning, in: NeurIPS Workshop on Human in the Loop Learning.

Gaillochet, M., Desrosiers, C., Lombaert, H., 2022. Taal: Test-time augmentation for active learning in medical image segmentation, in: Data Augmentation, Labeling, and Imperfections (MICCAI DALI).

Gal, Y., Ghahramani, Z., 2016. Dropout as a Bayesian approximation: representing model uncertainty in deep learning, in: Proceedings of the 33rd International Conference on International Conference on Machine Learning (ICML).

Gal, Y., Islam, R., Ghahramani, Z., 2017. Deep Bayesian Active Learning with Image Data, in: Proceedings of the 34th International Conference on Machine Learning (ICML).

Gao, M., Zhang, Z., Yu, G., Arik, S.O., Davis, L.S., Pfister, T., 2020. Consistency-Based Semi-supervised Active Learning: Towards Minimizing Labeling Cost, in: European Conference on Computer Vision (ECCV).

Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., He, K., 2018. Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour. arXiv:1706.02677 (accessed: 28 Jan. 2022) .

Hsu, W.N., Lin, H.T., 2015. Active Learning by Learning. Proceedings of the AAAI Conference on Artificial Intelligence 29.

Huang, S., Wang, T., Xiong, H., Huan, J., Dou, D., 2021. Semi-Supervised Active Learning with Temporal Output Discrepancy, in: IEEE International Conference on Computer Vision (ICCV).

Kim, K., Park, D., Kim, K.I., Chun, S.Y., 2021. Task-Aware Variational Adversarial Active Learning, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Kingma, D.P., Ba, J., 2015. Adam: A Method for Stochastic Optimization, in: 3rd International Conference for Learning Representations (ICLR).

Kirsch, A., van Amersfoort, J., Gal, Y., 2019. BatchBALD: Efficient and Diverse Batch Acquisition for Deep Bayesian Active Learning, in: Advances in Neural Information Processing Systems (NeurIPS).

Konyushkova, K., Sznitman, R., Fua, P., 2015. Introducing Geometry in Active Learning for Image Segmentation, in: IEEE International Conference on Computer Vision (ICCV), IEEE. pp. 2974–2982.

Konyushkova, K., Sznitman, R., Fua, P., 2019. Geometry in active learning for binary and multi-class image segmentation. Computer Vision and Image Understanding 182, 1–16.

Li, H., Yin, Z., 2020. Attention, Suggestion and Annotation: A Deep Active Learning Framework for Biomedical Image Segmentation, in: Medical Image Computing and Computer Assisted Intervention (MICCAI).

Li, X., Xia, M., Jiao, J., Zhou, S., Chang, C., Wang, Y., Guo, Y., 2023. HAL-IA: A Hybrid Active Learning framework using Interactive Annotation for medical image segmentation. Medical Image Analysis 88, 102862.

Litjens, G., Toth, R., van de Ven, W., Hoeks, C., Kerkstra, S., van Ginneken, B., Vincent, G., Guillard, G., Birbeck, N., Zhang, J., Strand, R., Malmberg, F., Ou, Y., Davatzikos, C., Kirschner, M., Jung, F., Yuan, J., Qiu, W., Gao, Q., Edwards, P.E., Maan, B., van der Heijden, F., Ghose, S., Mitra, J., Dowling, J., Barratt, D., Huisman, H., Madabhushi, A., 2014. Evaluation of prostate segmentation algorithms for MRI: The PROMISE12 challenge. Medical Image Analysis 18, 359–373.

Loshchilov, I., Hutter, F., 2017. SGDR: Stochastic Gradient Descent with Warm Restarts, in: International Conference on Learning Representations (ICLR).

Mahapatra, D., Bozorgtabar, B., Thiran, J.P., Reyes, M., 2018. Efficient Active Learning for Image Classification and Segmentation Using a Sample Selection and Conditional Generative Adversarial Network, in: Medical Image Computing and Computer Assisted Intervention (MICCAI).

Mittal, S., Tatarchenko, M., Çiçek, O., Brox, T., 2019. Parting with Illusions about Deep Active Learning. arXiv:1912.05361 .

Munjal, P., Hayat, N., Hayat, M., Sourati, J., Khan, S., 2022. Towards Robust and Reproducible Active Learning Using Neural Networks, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Nath, V., Yang, D., Landman, B.A., Xu, D., Roth, H.R., 2021. Diminishing Uncertainty Within the Training Pool: Active Learning for Medical Image Segmentation. IEEE Transactions on Medical Imaging 40, 2534–2547.

Nath, V., Yang, D., Roth, H.R., Xu, D., 2022. Warm Start Active Learning with Proxy Labels and Selection via Semi-supervised Fine-Tuning, in: Medical Image Computing and Computer Assisted Intervention (MICCAI).

Ozdemir, F., Peng, Z., Fuernstahl, P., Tanner, C., Goksel, O., 2021. Active learning for segmentation based on Bayesian sample queries. Knowledge-Based Systems 214, 106531.

Ozdemir, F., Peng, Z., Tanner, C., Fuernstahl, P., Goksel, O., 2018. Active Learning for Segmentation by Optimizing Content Information for Maximal Entropy, in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support.

Ren, P., Xiao, Y., Chang, X., Huang, P.Y., Li, Z., Gupta, B.B., Chen, X., Wang, X., 2021. A Survey of Deep Active Learning. ACM Computing Surveys 54, 180:1–180:40.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation, in: Medical Image Computing and Computer-Assisted Intervention (MICCAI).

Sener, O., Savarese, S., 2018. Active Learning for Convolutional Neural Networks: A Core-Set Approach, in: International Conference on Learning Representations (ICLR).

Settles, B., 2009. Active Learning Literature Survey. Technical Report. University of Wisconsin-Madison Department of Computer Sciences.

Shannon, C.E., 1948. A Mathematical Theory of Communication. Bell System Technical Journal 27, 379–423.

Sinha, S., Ebrahimi, S., Darrell, T., 2019. Variational Adversarial Active Learning, in: IEEE International Conference on Computer Vision (ICCV).

Smailagic, A., Costa, P., Young Noh, H., Walawalkar, D., Khandelwal, K., Galdran, A., Mirshekari, M., Fagert, J., Xu, S., Zhang, P., Campilho, A., 2018. MedAL: Accurate and Robust Deep Active Learning for Medical Image Analysis, in: 17th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 481–488.

Sourati, J., Gholipour, A., Dy, J.G., Kurugol, S., Warfield, S.K., 2018. Active Deep Learning with Fisher Information for Patch-Wise Semantic Segmentation, in: Stoyanov, D., Taylor, Z., Carneiro, G., Syeda-Mahmood, T., Martel, A., Maier-Hein, L., Tavares, J.M.R., Bradley, A., Papa, J.P., Belagiannis, V., Nascimento, J.C., Lu, Z., Conjeti, S., Moradi, M., Greenspan, H., Madabhushi, A. (Eds.), Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support (DLMIA). Springer International Publishing. volume 11045, pp. 83–91.

Sourati, J., Gholipour, A., Dy, J.G., Tomas-Fernandez, X., Kurugol, S., Warfield, S.K., 2019. Intelligent Labeling Based on Fisher Information for Medical Image Segmentation Using Deep Learning. IEEE Transactions on Medical Imaging 38, 2642–2653.

Top, A., Hamarneh, G., Abugharbieh, R., 2011. Active Learning for Interactive 3D Image Segmentation, in: Medical Image Computing and Computer-Assisted Intervention (MICCAI).

Wang, K., Zhang, D., Li, Y., Zhang, R., Lin, L., 2017. Cost-Effective Active Learning for Deep Image Classification. IEEE Transactions on Circuits and Systems for Video Technology 27, 2591–2600.

Xie, B., Yuan, L., Li, S., Liu, C.H., Cheng, X., 2022. Towards Fewer Annotations: Active Learning via Region Impurity and Prediction Uncertainty for Domain Adaptive Semantic Segmentation, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Yang, L., Zhang, Y., Chen, J., Zhang, S., Chen, D.Z., 2017. Suggestive Annotation: A Deep Active Learning Framework for Biomedical Image Segmentation, in: Medical Image Computing and Computer-Assisted Intervention (MICCAI).

Yoo, D., Kweon, I.S., 2019. Learning Loss for Active Learning, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Zhang, W., Zhu, L., Hallinan, J., Zhang, S., Makmur, A., Cai, Q., Ooi, B.C., 2022. BoostMIS: Boosting Medical Image Semi-Supervised Learning With Adaptive Pseudo Labeling and Informative Active Annotation, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Zhao, S., Song, J., Ermon, S., 2019. InfoVAE: Balancing Learning and Inference in Variational Autoencoders. Proceedings of the AAAI Conference on Artificial Intelligence 33, 5885–5892. Number: 01.