Task-Aware Active Learning for Endoscopic Polyp Segmentation

SUBMISSION DATE / POSTED DATE

12-05-2023 / 12-05-2023

CITATION

Thapa, Shrawan Kumar; Poudel, Pranav; Regmi, Sudarshan; Bhattarai, Binod; Stoyanov, Danail (2023): Task-Aware Active Learning for Endoscopic Polyp Segmentation. TechRxiv. Preprint. https://doi.org/10.36227/techrxiv.22810595.v1

DOI

10.36227/techrxiv.22810595.v1

Task-Aware Active Learning for Endoscopic Polyp Segmentation

Shrawan Kumar Thapa, Pranav Poudel, Sudarshan Regmi, Binod Bhattarai, Danail Stoyanov *Senior Member, IEEE*

***Abstract*— Semantic segmentation of polyps is one of the most important research problems in endoscopic image analysis. One of the main obstacles to researching such a problem is the lack of annotated data. Endoscopic annotations necessitate the specialist knowledge of expert endoscopists, and hence the difficulty of organizing arises along with tremendous costs in time and budget. To address this problem, we investigate an active learning paradigm to reduce the requirement of massive labeled training examples by selecting the most discriminative and diverse unlabeled examples for the task taken into consideration. To this end, we propose a task-aware active learning pipeline that considers not only the uncertainty that the current task model exhibits for a given unlabelled example but also the diversity in the composition of the acquired pool in the feature space of the model. We compare our method with the competitive baselines on two publicly available polyps segmentation benchmark datasets. Both qualitative and quantitative analysis show a significant improvement in performance when sampling on the semantic space of the model than image space, and also demonstrate complementary nature of using model uncertainty information. The code and implementation details are available at: `https://github.com/thetna/endo-active-learn`**

***Index Terms*— Active Learning, Computer Assisted Interventions, Endoscopic Image Analysis, Semantic Segmentation, Surgical AI**

I. Introduction

Polyp segmentation [1], [2] is a fundamental research problem in endoscopic image analysis. Automated polyp segmentation can help in the early diagnosis, detection, and treatment of colorectal disease by supporting endoscopists with computer-assisted detection and characterization systems. Such capabilities are needed to advance the toolkit available to endoscopists, enable standardization of adenoma detection rates, and potentially link to future robotic systems and automation [3]. The effectiveness of deep networks in learning the parameters for such tasks has already been demonstrated. However, most solutions demand a large number of training examples. Annotating such a large volume of endoscopic data needs domain experts, which incurs an immense cost in time and budget. Therefore, label-efficient methods are of utmost importance.

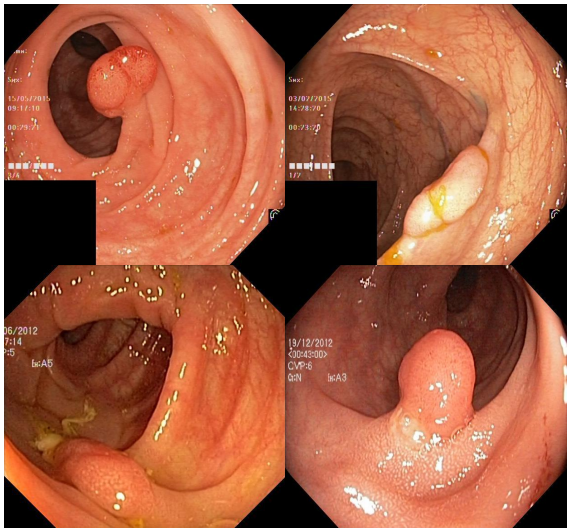


Fig. 1. Few randomly sampled examples from Kvasir-SEG [4].

Recently, to address the problem of annotated examples, several self-supervised learning algorithms are proposed [5] [6] [7] [8]. However, the performance of these approaches depends upon the overlap of the pre-text task with the downstream task. As a consequence, it demands a careful design of the pre-text task. Similarly, data augmentations with different geometric transformations, such as flipping and rotation, are another option to populate the training examples. These approaches do not effectively add true distribution variability. This is because the body organs such as Colons are tubular and rationally invariant. Figure 1 shows some of the images of the colons captured by the Endoscopes. Another viable option to augment the training data set is by generating realistic synthetic examples. Training large models such as Stable Diffusion [9] or similar for the medical domain demands

This work was supported by the Wellcome/EPSRC Centre for Interventional and Surgical Sciences (WEISS) [203145Z/16/Z]; Engineering and Physical Sciences Research Council (EPSRC) [EP/P027938/1, EP/R004080/1, EP/P012841/1]; The Royal Academy of Engineering Chair in Emerging Technologies scheme; and the EndoMapper project by Horizon 2020 FET (GA 863146).

SK Thapa and S. Regmi are with Nepal Applied Mathematics and Informatics Institute for research (NAAMII), Nepal (e-mail: {sk.thapa, sudarshan.regmi}@naamii.org.np)

P. Poudel is with Institute of Engineering (IOE), Pulchowk Campus, Nepal (email: 074BCT526.pranav@pcampus.edu.np)

B. Bhattarai is with the University of Aberdeen, UK (e-mail binod.bhattarai@abdn.ac.uk). He is the corresponding author of the paper.

D. Stoyanov is with University College London, UK (e-mail danail.stoyanov@ucl.ac.uk)

billions of annotated examples which is difficult to find.

Active Learning (AL) [10]–[12] has shown a lot of promise to become a viable solution to sub-sample the datasets by discarding redundant and less informative examples in computer vision. In AL, we repeatedly acquire labels using an acquisition function for a subset of an unlabeled set where the label acquisition is constrained by budget. Its task is to select the optimal subset of examples that enhance the model's performance when added to the training set. AL methods are emerging gradually in Biomedical Image Analysis [13]–[17]. NVIDIA's open-source platform MONAI [1] has launched an intelligent interactive data annotation tool called MONAI Label. Workshop with the theme "interpretable and label-efficient learning" [18] was organized in conjunction with MICCAI 2020. PathAL [19] and [20] are some of the recent works on Active Learning for Histopathology Image Analysis.

In active learning studies, acquisition functions fall mostly under three categories: 1. Uncertainty-based [21], [22], [23], [24] 2. Distribution-based [11], [10], and 3. Combining Uncertainty and distribution [20], [25]. Relying only on uncertainty as a selection criterion helps us to choose the examples from the region of the manifolds of the image where the model is less confident. However, it can not avoid selecting redundant images from the same manifold region, limiting the diversity. The distribution-based approaches address this issue by considering the selected samples' diversity. However, it is possible to miss the selection of difficult examples. So, the best bet is to combine the best of both worlds. Yet another difference that these two groups of method exhibits is that uncertainty-based methods are aware of downstream tasks. In contrast, the representative-based methods are task-agnostic.

Our contribution lies in developing a novel *task-aware* method for selecting both diverse and difficult examples for a downstream task and applying it to novel bio-medical image analysis tasks. To this end, we employ Coreset [26] sampling method on downstream *task-aware* feature space in our active learning pipeline for endoscopic polyp segmentation. Previous work on active learning [10] relying on Coreset based sampling was evaluated on classification problems. Unlike previous work, we also combine the uncertainty-based method to sample the unlabeled data. Our pipeline is as shown in Figure 2. There are three main components in our pipeline: a Learner (A), a Sampler (B), and an oracle (C). The learner is responsible for learning the parameters for a downstream task by using the examples whose labels are queried by the Oracle. We approximate the learner by UNet [27] which is one of the most widely used semantic segmentation networks for medical image analysis. However, networks other than this can also be employed without any difficulties. The sampler selects the examples and feeds them to the oracle to query the labels. As the data selected by the sampler directly influences the learner's performance, we argue the need for linkage between these two components. Therefore, we project all the data on the learner's feature space and apply the K-Center Greedy Algorithm similar to that in [10] to select the core-set examples from the unlabeled dataset that comprises a fraction of the total budget. This would help us to identify the diverse example on downstream task's feature space. Similarly, we acquire predictions for unlabeled examples and calculate uncertainty using the Best vs. Second Best Strategy (BvSB) [24], and obtain top uncertain samples. This would make the remaining fraction of the total budget. This approach helps us to focus on difficult examples. The sum of examples from both these approaches equals the total available budget. The trade-off is adjusted by empirical validation.

We summarize our contributions in the following points:

- We proposed a novel task-aware Coreset-based selection method in an active learning pipeline.
- We combine the Uncertainty based sampling technique with the task-aware Coreset-based sampling technique.
- We compare the proposed method with multiple task-agnostic approaches based on Coreset, and Variational Autoencoder on challenging data sets for Endoscopic polyp segmentation.
- We perform extensive quantitative and qualitative experiments to validate our approach.

The remaining paper is organized as follows. Section II covers some of the important works on active learning in biomedical image analysis. We present our method in details in Section III. Similarly, we discuss the experimental results and conclusion in Section IV and VI, respectively.

II. Related Works

A. Endoscopic Polyp Segmentation

The limited size of medical datasets is a well-known problem, and this hasn't eluded polyps segmentation tasks. ELKarazle, Khaled, et al. [28] have documented standard polyps datasets in their survey. The Kvasir-SEG [4] and the CVC-ClinicDB [29] used in this work have a total of 1000 and 612 annotated images, respectively. Similarly, the ETIS-Larib [30] and the CVC-ColonDB [31] have a mere 196 and 300 annotated samples, respectively. EndoTect[2] have 110k images, but only a thousand come with segmentation masks, and most of them are unlabeled. These examples show the difficulty in acquiring a large volume of binary masks for localizing the regions of polyps in endoscopic images. Several important works have been published on polyps segmentation, such as [32]–[35]. However, their focus has been primarily on architecture engineering. Works on sub-sampling the endoscopic data set to reduce annotation costs have been missing, and we believe our work provides a significant contribution in that regard.

B. Active Learning in Biomedical Image Analysis

PathAL [19] is one of the recently published works on an active learning framework for Histopathology image analysis. This work relies on uncertainty to the downstream task as a selection criterion. Examples with higher uncertainty are chosen to query their labels. Whilst, the examples with the lower uncertainty are assigned with the pseudo-labels predicted by the model of the task taken into consideration. This

[1]https:monai.io

[2]endotect.com

method also ignores diversity and ends up selecting redundant examples. Moreover, the examples on which the model is already confident may not add extra information to the model.

Recent work in medical image segmentation [36] employs the query-by-committee method to select the most informative samples. Stein Variational Gradient Descent method trains an ensemble of segmentation models. Then, an entropy-based uncertainty estimate is used to get an informativeness score for each unlabeled sample. The uncertainty estimate is combined with mutual information between pairs of labeled and unlabeled examples not to select the redundant examples. However, the diversity is maintained from the uncertain examples. Another work in segmentation [37] uses the Best versus Second Best Strategy to first select the most informative samples. Then, a distribution discrepancy measure, that employs cosine similarity between unlabeled and labeled sets, is used to select representative subsets from the selected informative samples. This method is yet another method to discard redundant examples from the most uncertain examples. From these methods, we can understand the importance of the selection function that gives diversity. However, the underlying idea of these methods is based on the uncertainty of the model for the downstream task. There is no explicit mechanism to identify the essential subset of examples for a given task from the whole data set. Hence, these methods also inherit the problem of other uncertainty-based methods.

IDEAL [20] uses a saliency map of unlabeled examples to train an auto-encoder to reconstruct the original map. The latent bottleneck feature is then used to cluster the images into k groups and train Random Forest Classifier online to rank samples based on its informativeness to train this classifier obtained using the AUC increment score from representative samples from each of the k clusters. However, the bottleneck features are not directly aware of the downstream task.

Yang et. al. [38] proposed an AL framework for gland and lymph node segmentation based on class conditional uncertainty as a criterion to select unlabelled data. The uncertainty is calculated based on variances of predictions from models trained using bootstrapping approach. Top K such examples are selected and a subset with $k < K$ examples, which are the most representative of features in the unlabeled set, is further sampled. Similarly, [39] uses a conditional generative adversarial network to generate synthetic examples and estimate the uncertainties to select data to query their labels.

Most of the above-mentioned methods modify the model training stages to induce biases of the active learning pipeline to the task model itself. Those have been proven to be effective as the works have shown, but the goal of this research is to propose such an AL pipeline, which makes the optimal utilization of information obtained from a model trained using a generic training paradigm. To this end, our method is inspired by the work of Shi et al. [25]. It presents an AL framework for skin lesion detection aiming to select both the difficult and the diverse examples. To this end, the paper proposed to do hashing on the image features computed in an unsupervised manner by applying Principal Component Analysis (PCA) [40] and clustering the images into different bins. Next, the paper proposed to sample the images from each bin uniformly. It also samples uncertain examples like our framework but uses the highest probability score of the classes. The lower the probability of the most certain class, the more uncertain the model is about the example. Nevertheless, the features computed in an unsupervised manner are not aware of the end task. Hence, the diverse examples on such sub-optimal features for the end task may not necessarily be diverse for the downstream task. Hence, we apply coreset [10] on latent space (output from final layer) of the encoder and we use BVSB strategy [24] as our uncertainty measure as the paper shows its superiority over entropy-based uncertainty measure.

In summary, most existing methods are either distribution-based or uncertainty based, and some use both. The problem with the distribution-based approach is being unaware of the end task, which makes the selection of data sub-optimal. Similarly, the uncertainty-based method inherits the property of the downstream task but fails to sample the diverse examples. Our pipeline brings the advantages from both approaches making the component of diversity task-aware.

III. METHOD

Active Learning is an iterative process to select a subset of examples (X^s) from a large pool of unlabeled set (X) to query their labels (Y). We label the examples (x, y) $\subset$ ($X \times Y$) incrementally and add to a set of the labeled examples (X^l). The labeled examples are used to train a network minimizing the objective of the end task ($\mathcal{L}$). Equation (1) summarises the Active Learning pipeline. Given any sampling function $\mathcal{A}$, the main goal of AL is to minimize the number of selection stages n to reduce the number of examples for which labels need to be queried.

$$\min_{n} \min_{\mathcal{L}} \mathcal{A}(\mathcal{L}(x, y; \theta) | X_0^s \subset \cdots \subset X_n^s \subset X). \tag{1}$$

To begin with annotation, we select the first batch X_0^s randomly, where subscript 0 denotes the first selection stage and superscript s indicates a selected set of examples to query their labels. Once oracle queries their labels, we add those examples to the pool of labeled examples $X^l = \{X_0^s \ U \ \emptyset\}$. These labeled examples act as seed annotations to guide the next selection stages. Fig. 2 depicts the proposed method. There are three major components in the pipeline A) Learner, B) Sampler, and C) Oracle. We discuss these in detail below.

A. Learner (A)

The role of a learner in the AL pipeline is to learn the parameters for a downstream task from the labeled set of examples. In our case, we are dealing with polyp segmentation. Thus, we choose U-Net [27], a widely-used semantic segmentation architecture for bio-medical image segmentation, to implement the learner. Suppose x represents an image with its corresponding ground-truth label y from labeled set X^l. When we feed in x to the model, the encoder projects the image into a low-dimensional vector, z. And the decoder reconstructs z back to the output $\hat{y}$, along with using different levels of

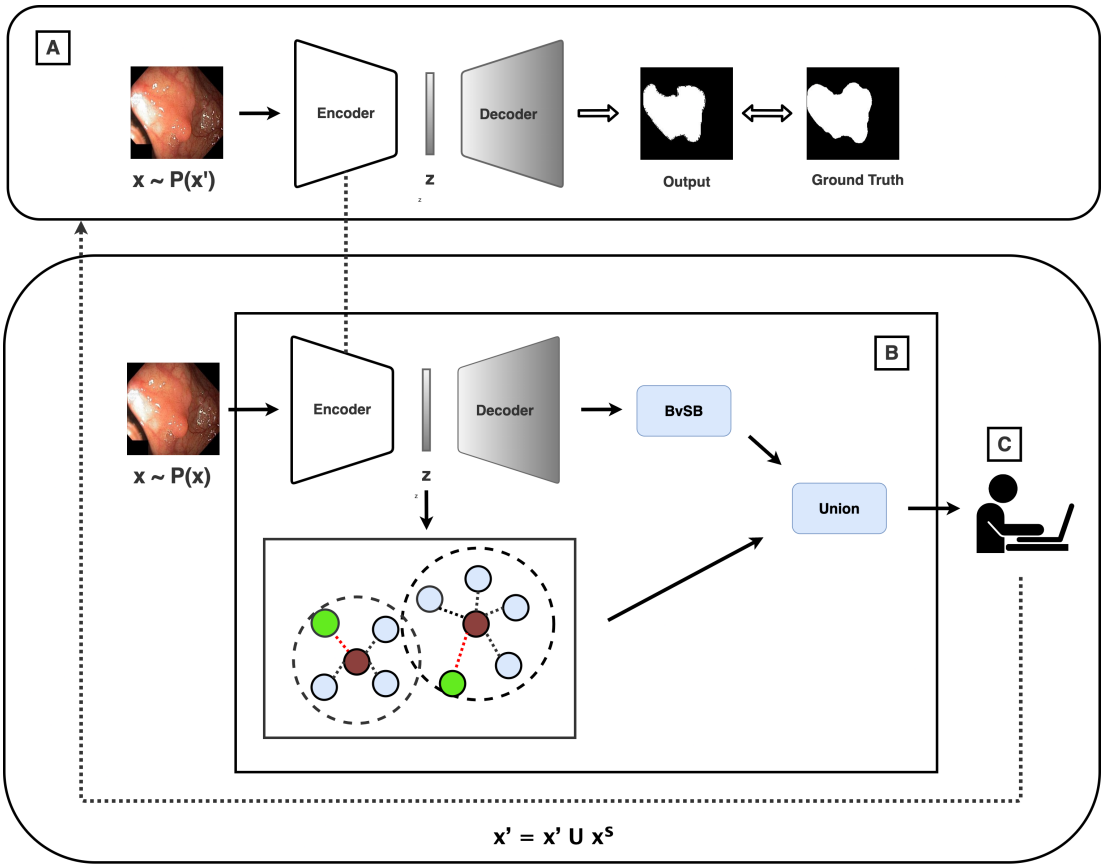


Fig. 2. This diagram shows the proposed Active Learning Pipeline. In this pipeline, (A) represents the model for the downstream task, also known as Learner. In our case, it is a semantic segmentation network. Similarly, (B) represents the sampler. The sampler shares the parameters of the learner, which makes our sampling technique task-aware. We extract the latent representations of the labelled and unlabelled data and employ Coreset and BvSB to sample both the task-aware diverse and uncertain examples. The selected examples are sent off to query their labels (c). The labelled examples and previously labelled data train the learner and re-iterate the process until the given budget.

features from the encoder. We minimize the objective given in (2) to train the network.

$$\mathcal{L}(y, \hat{y}) = \mathcal{L}_{\mathcal{CE}}(y, \hat{y}) + \mathcal{L}_{dice}(y, \hat{y}) \tag{2}$$

Here, $\mathcal{L}_{\mathcal{CE}}$ is a binary cross-entropy loss [41] and $\mathcal{L}_{dice}$ is dice loss [42]. Both are popular loss functions used in semantic segmentation.

Once we learn the parameters of U-Net from the available labeled examples, component B of the pipeline, the sampler, comes into play.

B. Sampler (B)

1) Uncertainty Based Sampling: In this stage, we feed unlabeled images to the model trained in the first stage and obtain the respective segmentation mask. Then, we use the BvSB method [24], which is one of the most competitive baselines to select the informative examples, to compute the uncertainty of the current model on given unlabeled images. This method uses the difference between the highest and the second-highest probability score predicted by the model. In the segmentation task, for each pixel (i, j), where $i \in H$, $j \in W$, the model predicts a categorical distribution denoted by a vector $\hat{y}(i, j) \in [0, 1]^C$, where C is the total number of distinct classes in the task. H and W are the height and the width of the input image/segmentation mask respectively. Then, the Best vs Second Best Score for each pixel is calculated as follows:

$$\mathcal{BvSB}(\hat{y}(i,j)) = 1 - [\max_{k \in \hat{y}(i,j)} \hat{y}(i,j) - \max_{l \in \hat{y}(i,j) \setminus k} \hat{y}(i,j)] \tag{3}$$

Lesser the difference in top-2 class prediction, the higher the uncertainty of the example for the model. Since polyps segmentation is a binary segmentation task, the score calculation for each image can be simplified as shown in the following equation:

$$\mathcal{BvSB}(\hat{y}) = 1 - \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} [\max(\hat{y}(i,j)) - \min(\hat{y}(i,j))] \tag{4}$$

We sample top B_u uncertain examples using the uncertainty score given by (4). The uncertainty is estimated from the predictions of the learner that models our downstream task; thus making this component task-aware.

2) Task-Aware Core-set: Furthermore, we sample B_d diverse examples. We use the parameters of the encoder learned in stage A to project both labeled and unlabeled images into a latent space, Z. The procedure is carried out by first getting the bottleneck feature from the encoder of dimension

$H/16 \times W/16 \times 512$. We, then, perform global average pooling and flattening operations to get the latent image feature vector $z \in \mathbb{R}^{512}$. Since we train the learner to minimize the objective of the downstream task, the latent representations of the images are optimized for the same. Next, we propose to select a core-set [26] of the dataset based on this latent space to query their labels. To compute the coreset, we employ the K-Center-Greedy algorithm on this latent space as shown in Algorithm 1.

Algorithm 1 k-Center-Greedy

Input: latent representation of data $\mathbf{Z}$, existing pool $\mathbf{s}^0$ and a budget b
Initialize $\mathbf{s} = \mathbf{s}^0$
repeat
  $u = \arg\max_{i \in [n] \setminus s} \min_{j \in s} \Delta(\mathbf{z}_i, \mathbf{z}_j)$
  $\mathbf{s} = \mathbf{s} \cup \{u\}$
until $|\mathbf{s}| = b + |\mathbf{s}^0|$
return $\mathbf{s} \setminus \mathbf{s}^0$

$\Delta(\mathbf{z}_i, \mathbf{z}_j)$ in algorithm 1 measures distance between two latent features. We used euclidean distance as a distance metric.

This selection process for diversity is also depicted visually in the form of a graph in Fig. 2 block B (lower sub-block). Each node in the graph represents an image, and the node's feature is initialized with the latent features extracted from the learner's encoder. Sky-blue nodes denote unlabeled examples, red nodes denote the labeled examples, and green nodes denote the examples selected to query their labels in the current stage. After the completion of the selection process, selected examples become a part of labeled examples. The edges between the nodes represent the euclidean distances, where the length is proportional to the magnitude of the distances. Here, we create a territory of the nearest examples for each red node and select the farthest node amongst these as illustrated in the figure. Examples with the least euclidean distances are likely to be duplicates of the selected ones and provide redundant information to the downstream task. Hence, this selection strategy helps us to obtain a subset of representative examples of the dataset by discarding the redundant ones for the end task. Since the core-set we obtain is aware of the downstream task, we term our approach as *Task-Aware Coreset (TA-Coreset)*.

3) Combining TA-Coreset with Uncertainty: Here, B_u and B_d are functions of γ such that $B_u = \gamma * B$ and $B_d = (1 - \gamma) * B$, where B is the total number of examples at a selection stage. We obtain X^s from the union of uncertain and diverse sets having B_u and B_d examples in each set respectively. It should be noted that both sets are disjoint. Indeed, it is done to show the complementary nature of these methods. If in case, the same example(s) is selected by both methods, the union of the two sets doesn't result in a total of B samples. In such cases, we continue the same sampling scheme for the remaining number to be sampled from the budget. Before proceeding to the next cycle, we add the selected examples in the current cycle to the existing pool of labeled examples. This iterative sampling process is summarized in algorithm 2. More discussion of this is included in section V.

Algorithm 2 Combining TA-Coreset with Uncertainty

Input: latent representation of data $\mathbf{Z}$, existing pool $\mathbf{s}^0$, unlabelled set X^U, a budget B, sampling ratio γ
Initialize $\mathbf{s} = \phi$
repeat
  $B_i = B - |\mathbf{s}|$
  $B_u = \gamma * B_i$
  $B_d = (1 - \gamma) * B_i$
  $\mathbf{s_u} = \{B_u$ samples from X^U using equation 4$\}$
  $\mathbf{s_d} = \{B_d$ samples from $X^U \cup \mathbf{s}^0$ using algorithm 1$\}$
  $\mathbf{s} = \mathbf{s} \cup \mathbf{s_c} \cup \mathbf{s_d}$
  $\mathbf{s}^0 = \mathbf{s}^0 \cup \mathbf{s}$
  $X^U = X^U \setminus \mathbf{s}^0$
until $|\mathbf{s}| = B$
return $\mathbf{s}$

C. Oracle (C)

We query the labels of the selected set, X^s from the Oracle. After retrieving their label, the selected set is appended to the labelled set ($X^l = X^l \cup X^s$), and the selected set ($X^s = \emptyset$) is emptied. This cycle is repeated till the budget limit is reached.

IV. Experiments

This section presents the details of the experiments we performed to validate our hypothesis. We start with a brief description of the datasets, baselines, and evaluation methods, followed by both quantitative and qualitative evaluations.

A. Datasets

We perform extensive experiments on Kvasir-SEG [4] and Clinic-DB [29]. Fig. 3 shows a pair of randomly selected images from each dataset. These are two important benchmark datasets publicly available for polyp segmentation.

Kvasir-SEG consists of 1,000 colonoscopy images with polyp masks. We used 900 of them for training and the rest for validation. We reported our performance on a smaller test dataset provided by the same project, identified as sessile-Kvasir-SEG consisting of 196 images. Clinic-DB is a similar dataset, but data points amounting only to 612 images. We randomly selected 112 images as a test set and 100 from the remaining 500 images as a validation set. The validation set is used to select the best model during training. The model is then evaluated on the respective test sets.

B. Baselines:

We compared our method with a wide range of competitive baselines. They were selected to prove or disprove our hypothesis. We hypothesized that the features need to be task-aware, and uncertainty information is complementary for further performance improvement.

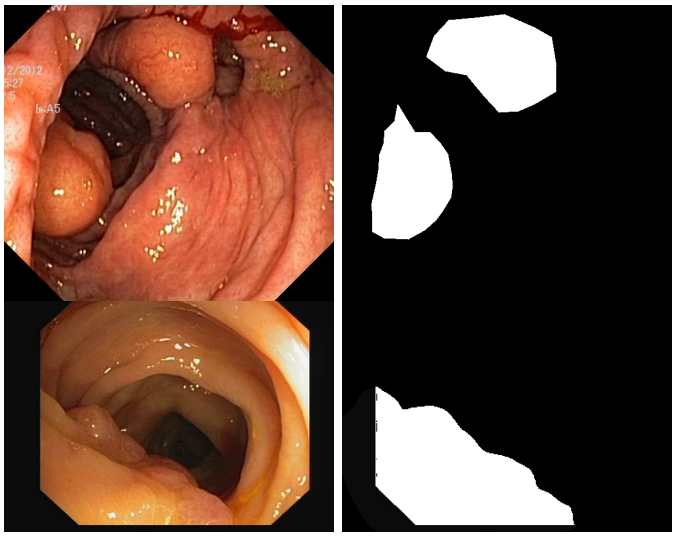


Fig. 3. Two upper images are from Kvasir-SEG, and the lower images are from CVC-ClinicDB data set.

Random is the most commonly used technique to sub-sample the training examples. We applied **Principal Component Analysis (PCA)** [40], and compressed the images to the dimension of 512, which is equal to that of the latent representations in our method. We applied Coreset [10] on PCA-compressed features, which we denote as **PCA-Coreset**. **Uncertainty** [24] is another sampling technique to find the most informative examples. Finally, we compared our performance with **VAAL** [11], one of the most popular task-agnostic active learning methods. Even though our method is not task-agnostic, we made a comparison with VAAL to shed light on the importance of task-aware active learning. However, we have intentionally avoided comparing our method to promising works such as PathAL [19] and Diminishing Uncertainty [36] because these methods induce biases of AL sampling process to task model training as well. Our objective is not to make any interventions to task model training, thus minimizing complexity. Additionally, we aim to make the sampling process independent so that the already trained model can also utilize the proposed data acquisition process to improve its performance in a label-efficient way.

C. Evaluation metrics:

We report the performance on both data sets using the metric of Intersection Over Union (mIOU) at *different selection stages*. We ran the experiment five times and report its mean and standard deviation.

$$\text{IoU} = \frac{\text{true positive}}{\text{true positive + false positive + false negative}}$$

In addition to this, we present extensive qualitative analysis to validate the proposed method.

D. Implementation Details:

We trained U-Net as our task model for polyps segmentation. U-Net is a popular and widely used deep architecture for medical image segmentation. Since our method is modular, any other architecture can be plugged easily into the pipeline. Please note that our contribution is on engineering the data by selecting the most discriminative examples. Thus, engineering the semantic segmentation architecture is beyond the scope of this work. We used Adam optimizer [43] ($\beta 1 = 0.9, \beta 2 = 0.999$) with a learning rate of 2×10^{-4}. We trained the model for 100 epochs in each cycle with a batch size of 8. We resized the image into the dimension of 256×256. We then divided the pixel values by 255 to get them in the range of 0-1, followed by normalization with a mean and standard deviation of 0.5.

For AL experimentation, we initialized our labelled pool with randomly selected 100 examples for Kvasir-SEG dataset, simultaneously keeping the sampling budget size of 100. In the case of Clinic-DB dataset, we initialized with randomly selected 40 examples keeping the same sampling budget size.

E. Quantitative Evaluations

1) Kvasir-SEG:: Tables I and II summarise the performance comparison on the Kvasir-SEG dataset with the baselines. The results are the mean and standard deviation of mean IOU from five different trials with each trial initialized with different seeds. Numerals representing each column indicate the performance result from model trained with n labeled examples. For example, 200 represents results from the model trained by adding to the current labeled set the examples sampled at the first selection stage by the respective AL pipeline. We have excluded the performance of the models from the first cycle of training (training set size of 100) since they were trained by initializing using the same random samples. From table I we can observe that using task-aware features to select diverse examples helps improve performance from respective task-independent counterparts. TA-Coreset outperforms other baseline methods most of the time. Also, when it is comparatively underperforming in terms of mean performance across five trials, it is still within the range of standard deviation of the best-performing model at that stage. From the metrics, we can also verify that using k-center algorithm on task-aware features significantly boosts performance from using it on the image space features extracted using PCA.

2) CVC-ClinicDB:: Tables III and IV summarise the performance comparison on the CVC-ClincDB dataset with the baselines. The results are the mean and standard deviation of mean IOU from five different trials with each trial initialized with different seeds. All the other conventions are similar to that in section A. We have excluded the performance of the model from the first cycle in this experiment as well following the same reasoning. From table III, we can similarly observe that task-aware features proved to be more significant than the task-independent sampling, though, in this dataset, TA-Coreset is competing closer to PCA-derived features, which is a deviation from that on Kvasir.

3) Combining Uncertainty and TA-Coreset: Tuning γ: As our method is based on diversity sampling, we performed experiments on combining samples from the Uncertainty based acquisition function and TA-Coreset. We uniformly sampled the values of γ from 0 to 1, which corresponds to the fraction

TABLE I
PERFORMANCE COMPARISON ON KVASIR-SEG DATASET: COMPARISON ON IMAGE FEATURES

Method	Mean IOU				
	200	300	400	500	600
Random	68.14 ± 0.40	73.67 ± 0.58	77.48 ± 0.69	81.63 ± 0.86	85.61 ± 0.58
VAAL [11]	$\mathbf{68.38 \pm 0.46}$	73.58 ± 1.21	$\mathbf{78.41 \pm 1.06}$	81.82 ± 1.08	85.59 ± 0.52
PCA-Coreset	64.71 ± 1.02	70.20 ± 1.01	75.43 ± 0.95	78.48 ± 0.58	81.96 ± 0.56
TA-Coreset (ours)	67.99 ± 0.97	$\mathbf{74.02 \pm 1.13}$	78.28 ± 0.85	$\mathbf{81.92 \pm 1.08}$	$\mathbf{86.89 \pm 0.65}$

TABLE II
PERFORMANCE COMPARISON ON KVASIR-SEG DATASET: COMBINATION OF UNCERTAINTY AND IMAGE FEATURES

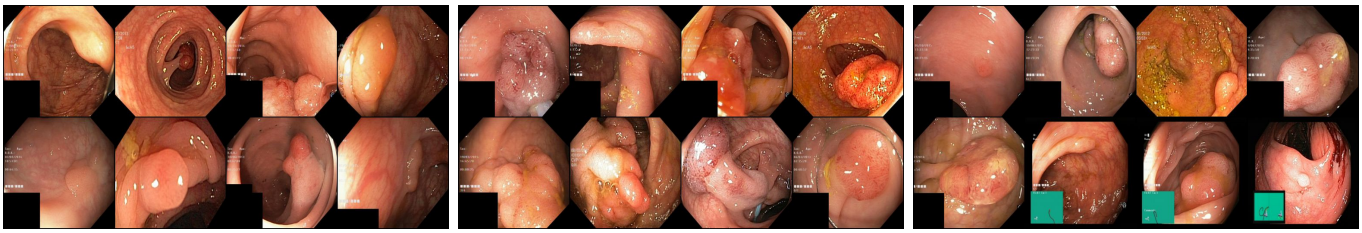
Method	Mean IOU				
	200	300	400	500	600
Uncertainty [24]	66.69 ± 0.56	73.96 ± 0.75	81.08 ± 1.14	$\mathbf{85.97 \pm 0.57}$	89.03 ± 0.18
Uncertainty + PCA [25]	65.92 ± 0.53	71.73 ± 0.44	77.73 ± 0.78	83.46 ± 0.40	87.79 ± 0.41
Uncertainty + TA-Coreset (ours)	$\mathbf{67.93 \pm 0.53}$	$\mathbf{75.35 \pm 1.00}$	$\mathbf{82.68 \pm 0.80}$	85.40 ± 0.53	$\mathbf{89.24 \pm 0.41}$

TABLE III
PERFORMANCE COMPARISON ON CVC-CLINICDB DATASET: COMPARISON ON IMAGE FEATURES

Method	Mean IOU				
	80	120	160	200	240
Random	78.92 ± 0.51	82.38 ± 0.35	84.83 ± 0.90	85.82 ± 0.41	86.77 ± 0.62
VAAL [11]	79.02 ± 0.92	82.42 ± 0.42	84.40 ± 0.73	85.21 ± 0.38	87.69 ± 0.29
PCA-Coreset	79.65 ± 0.69	83.67 ± 0.28	86.22 ± 0.29	87.47 ± 0.26	88.42 ± 0.29
TA-Coreset (Ours)	$\mathbf{79.94 \pm 0.29}$	$\mathbf{84.33 \pm 0.44}$	$\mathbf{87.07 \pm 0.40}$	$\mathbf{88.14 \pm 0.25}$	$\mathbf{88.78 \pm 0.16}$

TABLE IV
PERFORMANCE COMPARISON ON CVC-CLINICDB DATASET: COMBINATION OF UNCERTAINTY AND IMAGE FEATURES

Method	Mean IOU				
	80	120	160	200	240
Uncertainty [24]	79.18 ± 0.36	84.01 ± 0.27	85.96 ± 0.29	87.93 ± 0.23	88.94 ± 0.30
Uncertainty + PCA [25]	$\mathbf{80.25 \pm 0.36}$	83.46 ± 0.45	87.05 ± 0.33	87.78 ± 0.23	$\mathbf{88.98 \pm 0.35}$
Uncertainty + TA-Coreset (ours)	79.33 ± 0.64	$\mathbf{84.60 \pm 0.40}$	$\mathbf{87.12 \pm 0.19}$	$\mathbf{88.02 \pm 0.28}$	88.61 ± 0.13



A) B) C)

Fig. 4. Examples sampled by Uncertainty+TA-Coreset method in third stage of selection which are A) also sampled by Uncertainty-based method, but not Coreset B) also sampled by Coreset, but not Uncertainty-based method C) not sampled by both (unique to the combination)

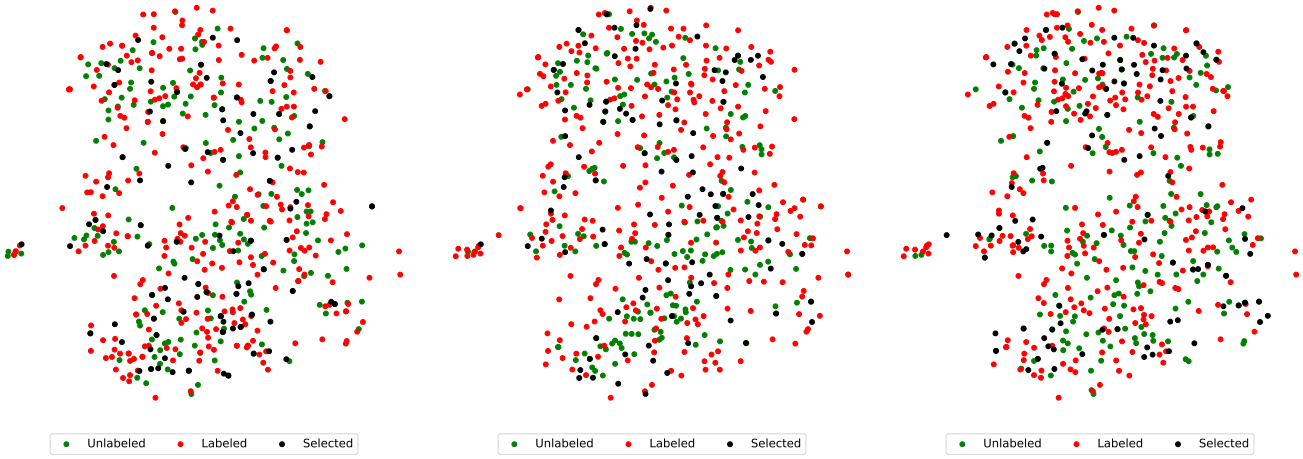


Fig. 5. T-SNE plots showing the comparison of selection of unlabelled examples at the third selection stage on Kvasir-SEG. Left, middle, and right plots show selection by Random, PCA-Coreset, and TA-Coreset, respectively (Zoom in for the better view).

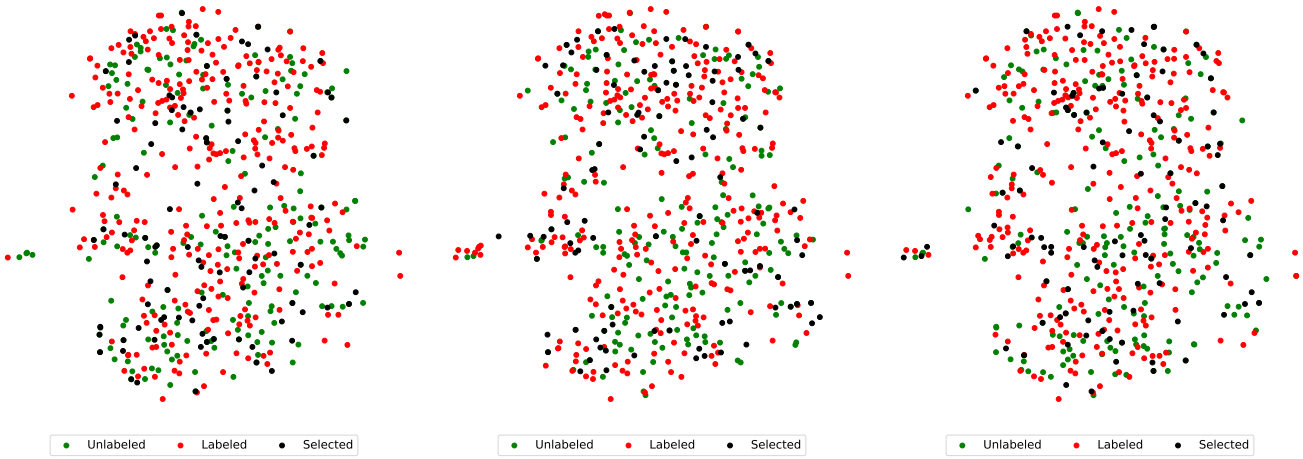


Fig. 6. T-SNE plots showing the comparison of selection of unlabelled examples at the third selection stage on Kvasir-SEG. Left, middle, and right plots show selection by Uncertainty, TA-Coreset, and Uncertainty+TA-Coreset, respectively (Zoom in for the better view).

of uncertain examples to sample from the total budget. When the value of $\gamma = 0$, this is equivalent to TA-Coreset, and when $\gamma = 1$, this matches with Uncertainty. The ablation studies on Kvasir-SEG [4] are shown in fig. 7 (left), and test results on CVC-ClinicDB [29] is shown in fig.7 (right). For both datasets, we observed the most optimal performance when $\gamma = 0.5$. This also demonstrates that our method is complementary to the existing uncertainty-based method.

Table II shows the effect of adding examples considering the distribution of unlabeled images to the already selected uncertain set. There is a boost in performance, especially in earlier stages. The competitive performance in later stages is because as the unlabeled pool starts shrinking, the number of unique examples providing new information could also shrink. Additionally, It should be noted that this approach isn't only complementary to uncertainty information but also provides a significant boost over diverse sampling from image/feature space (Table I and Table II). The metrics clearly show significant improvement from both PCA and TA-Coreset, specifically deviating in performance towards later stages.

In the case of combination in table IV, the performance came out much tighter in this dataset with other baseline methods. This could be explained by the much smaller set we started with (a total of only 400 images as opposed to 900 in Kvasir) which is similar to the observation in later stages of the Kvasir dataset.

V. Qualitative Evaluations

Fig. 4 illustrates several examples (Kvasir-SEG dataset) sampled by each method in the third selection stage. Fig. 4A) are common examples sampled by BvSB and combination but not by coreset, 4B) are sampled by coreset and combination, but not by BvSB, and 4C) are unique to combination only. In fig. 5 and fig.6, we summarize the selection behaviour of different sampling techniques in Kvasir-SEG data set with help of the tSNE [44] plots. In the diagram, the green dots are unlabeled examples, the red dots are labeled examples from the previous selection stage, and the black dot represents the examples selected in the current stage of annotation. The black dots were green before selection. The unlabeled set has been reduced using random sampling for clarity of the plot.

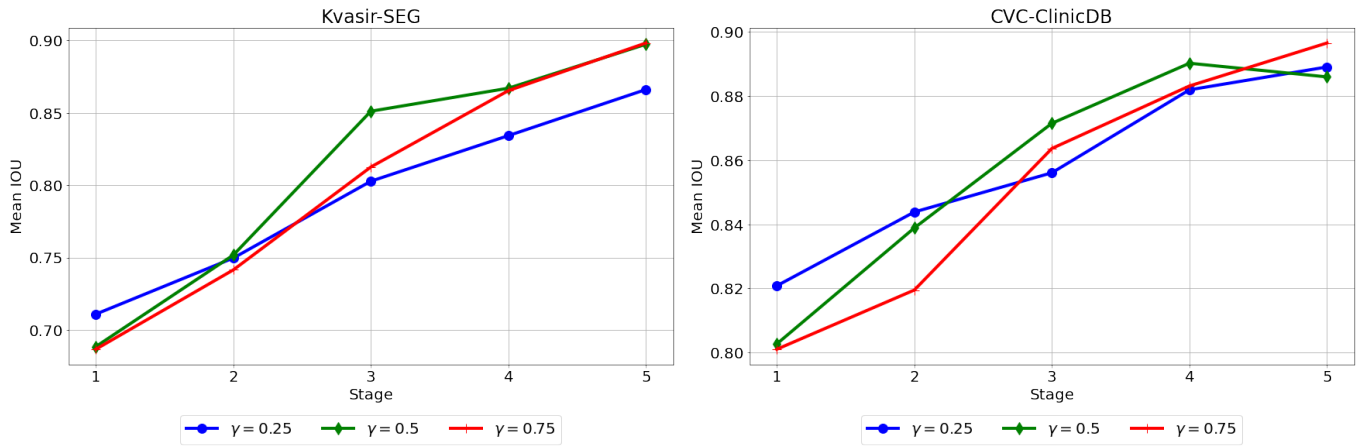


Fig. 7. Performance study on different values of γ for combining Uncertainty and TA-Coreset $(1-\gamma)$ on Kvasir (left) and CVC (right). When $\gamma = 0$, it is equivalent to TA-Coreset. We uniformly vary the weight from 0.25 to 0.75. This graph shows that our method is complementary to Uncertainty based Active Learning methods.

1) Task Aware vs Task Agnostic Features: Fig. 5 demonstrates the importance of task-aware features for the selection of an optimal set. From these plots, we can observe that random (left) selects the examples uniformly throughout the manifold. Similarly, PCA-coreset (middle) also covers the whole manifold, though the selected examples are more spread out than in random sampling. This shows the efficacy of Coreset in sampling the unlabeled set considering the distribution of input data. In contrast, TA-Coreset (right) concentrates on certain regions of the image manifold. The selected examples are more crowded in the upper and lower regions. The middle region looks sparse as compared to the selection in the other two methods. These observations depict that even though sampling in task-agnostic image space looks representative of the dataset, sampling in task-aware latent representation proved to be more significant in terms of selecting for maximizing model performance.

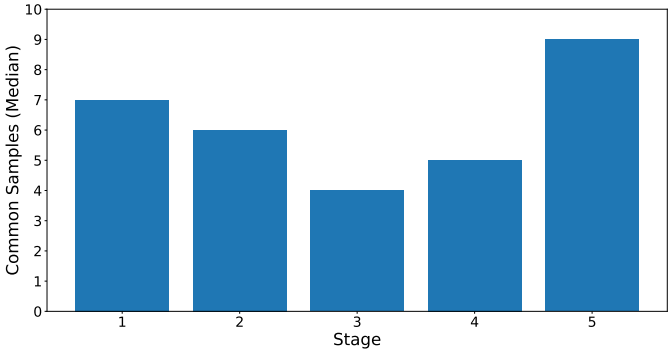


Fig. 8. Median Number of Common Examples sampled by TA-Coreset and Uncertainty at each stage of sampling

2) Importance of Combination: Fig. 6 demonstrates the importance of combining uncertainty and diversity for the selection of an optimal set. From these plots, we can see that TA-Coreset (Middle) misses the uncertain/informative example, as recognized by the uncertainty-based sampling method (Left), from the lower-middle part of the image space. When their sampling is combined (Right), we can see it covers regions with both uncertain examples and diverse task-aware features. This complementary nature is also recorded in fig. 8. The figure shows the median number of common examples sampled by the uncertainty-based method and TA-Coreset-based method at each stage of the active learning cycle. Since the budget was set to 100, and each method was asked to sample 50% ($\gamma = 0.5$) of the budget, we can observe that only a few examples are common selections from both methods. Similarly, From simple perception in fig. 4, we can observe that they are very different examples. This shows that our method can sample examples identified by BvSB and Coreset if they were used on their own without combination, and also its distinct samples.

VI. Conclusions

In this paper, we present a novel task-aware active learning framework for endoscopic image analysis. We combined diversity sampling on task-aware feature space with uncertainty information from the task model. We employed the proposed method on the polyp segmentation task and tested it on two publicly available datasets. We observed a superior performance from the extensive experiments compared to the multiple competitive baselines, validating the hypothesis that the feature required for sampling coreset of the dataset should be task aware. Furthermore, we also noted that the addition of model uncertainty information proved to be complementary though the performance starts getting competitive with availability of a smaller pool of unlabeled set. Though the value of γ was the same for both datasets in our experiments, it should be noted that the hyperparameter was determined by tuning the pipeline for each dataset. In future work, we will extend our pipeline for multi-tasking framework in Endoscopic image analysis.

References

[1] P. Brandao, E. Mazomenos, G. Ciuti, R. Caliò, F. Bianchi, A. Menciassi, P. Dario, A. Koulaouzidis, A. Arezzo, and D. Stoyanov, "Fully convolutional neural networks for polyp segmentation in colonoscopy," in

Medical Imaging 2017: Computer-Aided Diagnosis, vol. 10134. SPIE, 2017, pp. 101–107.

[2] S. Ali, M. Dmitrieva, N. Ghatwary, S. Bano, G. Polat, A. Temizel, A. Krenzer, A. Hekalo, Y. B. Guo, B. Matuszewski *et al.*, "Deep learning for detection and segmentation of artefact and disease instances in gastrointestinal endoscopy," *Medical image analysis*, vol. 70, p. 102002, 2021.

[3] O. F. Ahmad, Y. Mori, M. Misawa, S.-e. Kudo, J. T. Anderson, J. Bernal, T. M. Berzin, R. Bisschops, M. F. Byrne, P.-J. Chen *et al.*, "Establishing key research questions for the implementation of artificial intelligence in colonoscopy: a modified delphi method," *Endoscopy*, vol. 53, no. 09, pp. 893–901, 2021.

[4] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. d. Lange, D. Johansen, and H. D. Johansen, "Kvasir-seg: A segmented polyp dataset," in *International Conference on Multimedia Modeling*. Springer, 2020, pp. 451–462.

[5] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *NeurIPS*, 2017.

[6] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.

[7] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar *et al.*, "Bootstrap your own latent-a new approach to self-supervised learning," *Advances in neural information processing systems*, vol. 33, pp. 21 271–21 284, 2020.

[8] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[9] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," 2021.

[10] O. Sener and S. Savarese, "Active learning for convolutional neural networks: A core-set approach," in *International Conference on Learning Representations*, 2018.

[11] S. Sinha, S. Ebrahimi, and T. Darrell, "Variational adversarial active learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5972–5981.

[12] R. Caramalau, B. Bhattarai, and T.-K. Kim, "Sequential graph convolutional network for active learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9583–9592.

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

[14] C.-T. Li, H.-W. Tsai, T.-L. Yang, J.-C. Lin, N.-H. Chow, Y. H. Hu, K.-S. Cheng, and P.-C. Chung, "Imbalance-effective active learning in nucleus, lymphocyte and plasma cell detection," in *Interpretable and Annotation-Efficient Learning for Medical Image Computing*. Springer, 2020, pp. 223–232.

[15] M. Bernhardt, D. C. Castro, R. Tanno, A. Schwaighofer, K. C. Tezcan, M. Monteiro, S. Bannur, M. P. Lungren, A. Nori, B. Glocker *et al.*, "Active label cleaning for improved dataset quality under resource constraints," *Nature communications*, vol. 13, no. 1, pp. 1–11, 2022.

[16] J. Wu, S. Ruan, C. Lian, S. Mutic, M. A. Anastasio, and H. Li, "Active learning with noise modeling for medical image annotation," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE, 2018, pp. 298–301.

[17] S. C. Hoi, R. Jin, J. Zhu, and M. R. Lyu, "Batch mode active learning and its application to medical image classification," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 417–424.

[18] J. Cardoso, H. Van Nguyen, N. Heller, P. H. Abreu, I. Isgum, W. Silva, R. Cruz, J. P. Amorim, V. Patel, B. Roysam *et al.*, *Interpretable and Annotation-Efficient Learning for Medical Image Computing: Third International Workshop, iMIMIC 2020, Second International Workshop, MIL3ID 2020, and 5th International Workshop, LABELS 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4–8, 2020, Proceedings*. Springer Nature, 2020, vol. 12446.

[19] W. Li, J. Li, Z. Wang, J. Polson, A. E. Sisk, D. P. Sajed, W. Speier, and C. W. Arnold, "Pathal: An active learning framework for histopathology image analysis," *IEEE Transactions on Medical Imaging*, 2021.

[20] D. Mahapatra, A. Poellinger, L. Shao, and M. Reyes, "Interpretability-driven sample selection using self supervised learning for disease classification and segmentation," *IEEE transactions on medical imaging*, vol. 40, no. 10, pp. 2548–2562, 2021.

[21] Y. Gal, R. Islam, and Z. Ghahramani, "Deep bayesian active learning with image data," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1183–1192.

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

[23] D. Yoo and I. S. Kweon, "Learning loss for active learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 93–102.

[24] A. J. Joshi, F. Porikli, and N. Papanikolopoulos, "Multi-class active learning for image classification," in *2009 ieee conference on computer vision and pattern recognition*. IEEE, 2009, pp. 2372–2379.

[25] X. Shi, Q. Dou, C. Xue, J. Qin, H. Chen, and P.-A. Heng, "An active learning approach for reducing annotation cost in skin lesion analysis," in *International Workshop on Machine Learning in Medical Imaging*. Springer, 2019, pp. 628–636.

[26] P. K. Agarwal, S. Har-Peled, and K. R. Varadarajan, "Geometric approximation via coresets survey," *Current Trends in Combinatorial and Computational Geometry, E. Welzl, ed., Cambridge University Press, Cambridge*, 2006.

[27] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[28] K. ELKarazle, V. Raman, P. Then, and C. Chua, "Detection of colorectal polyps from colonoscopy using machine learning: A survey on modern techniques," *Sensors*, vol. 23, no. 3, p. 1225, 2023.

[29] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilariño, "Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," *Computerized medical imaging and graphics*, vol. 43, pp. 99–111, 2015.

[30] J. Silva, A. Histace, O. Romain, X. Dray, and B. Granado, "Towards embedded polyp detection in wce image," *International Journal of Computer Assisted Radiology and Surgery*, vol. 9, no. 2, pp. 283–293, 2014.

[31] J. Bernal, J. Sánchez, and F. Vilarino, "Towards automatic polyp detection with a polyp appearance model," *Pattern Recognition*, vol. 45, no. 9, pp. 3166–3182, 2012.

[32] D. Jha, P. H. Smedsrud, D. Johansen, T. de Lange, H. D. Johansen, P. Halvorsen, and M. A. Riegler, "A comprehensive study on colorectal polyp segmentation with resunet++, conditional random field and test-time augmentation," *IEEE journal of biomedical and health informatics*, vol. 25, no. 6, pp. 2029–2040, 2021.

[33] T. Kim, H. Lee, and D. Kim, "Uacanet: Uncertainty augmented context attention for polyp segmentation," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 2167–2175.

[34] N. T. Duc, N. T. Oanh, N. T. Thuy, T. M. Triet, and V. S. Dinh, "Colonformer: an efficient transformer based method for colon polyp segmentation," *IEEE Access*, vol. 10, pp. 80 575–80 586, 2022.

[35] J. Kang and J. Gwak, "Ensemble of instance segmentation models for polyp segmentation in colonoscopy images," *IEEE Access*, vol. 7, pp. 26 440–26 447, 2019.

[36] V. Nath, D. Yang, B. A. Landman, D. Xu, and H. R. Roth, "Diminishing uncertainty within the training pool: Active learning for medical image segmentation," *IEEE Transactions on Medical Imaging*, vol. 40, no. 10, pp. 2534–2547, 2020.

[37] H. Li and Z. Yin, "Attention, suggestion and annotation: a deep active learning framework for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 3–13.

[38] L. Yang, Y. Zhang, J. Chen, S. Zhang, and D. Z. Chen, "Suggestive annotation: A deep active learning framework for biomedical image segmentation," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2017, pp. 399–407.

[39] D. Mahapatra, B. Bozorgtabar, J.-P. Thiran, and M. Reyes, "Efficient active learning for image classification and segmentation using a sample selection and conditional generative adversarial network," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 580–588.

[40] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and intelligent laboratory systems*, vol. 2, no. 1-3, pp. 37–52, 1987.

[41] M. Yi-de, L. Qing, and Q. Zhi-Bai, "Automated image segmentation using improved pcnn model based on cross-entropy," in *Proceedings of 2004 International Symposium on Intelligent Multimedia, Video and Speech Processing, 2004.* IEEE, 2004, pp. 743–746.

[42] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. Jorge Cardoso, "Generalised dice overlap as a deep learning loss function for highly

unbalanced segmentations," in *Deep learning in medical image analysis and multimodal learning for clinical decision support*. Springer, 2017, pp. 240–248.

[43] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.

[44] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.