Predictive Accuracy-Based Active Learning for Medical Image Segmentation

Jun Shi
Shulan Ruan
Ziqi Zhu
Minfan Zhao
Hong An
Xudong Xue
Bing Yan

School of Computer Science and Technology, University of Science and Technology of China
School of Data Science, University of Science and Technology of China
Laoshan Laboratory
Hubei Cancer Hospital, Tongji Medical College, Huazhong University of Science and Technology
Department of Radiation Oncology, The First Affiliated Hospital of USTC, Division of Life Sciences and Medicine, University of Science and Technology of China

Abstract

Active learning is considered a viable solution to alleviate the contradiction between the high dependency of deep learning-based segmentation methods on annotated data and the expensive pixel-level annotation cost of medical images. However, most existing methods suffer from unreliable uncertainty assessment and the struggle to balance diversity and informativeness, leading to poor performance in segmentation tasks. In response, we propose an efficient Predictive Accuracy-based Active Learning (PAAL) method for medical image segmentation, first introducing predictive accuracy to define uncertainty. Specifically, PAAL mainly consists of an Accuracy Predictor (AP) and a Weighted Polling Strategy (WPS). The former is an attached learnable module that can accurately predict the segmentation accuracy of unlabeled samples relative to the target model with the predicted posterior probability. The latter provides an efficient hybrid querying scheme by combining predicted accuracy and feature representation, aiming to ensure the uncertainty and diversity of the acquired samples. Extensive experiment results on multiple datasets demonstrate the superiority of PAAL. PAAL achieves comparable accuracy to fully annotated data while reducing annotation costs by approximately 50% to 80%, showcasing significant potential in clinical applications. The code is available at https://github.com/shijun18/PAAL-MedSeg.

1 Introduction

Recently, supervised deep learning methods have been widely applied to medical image segmentation tasks, such as delineating organs and lesions [Wang et al., 2022]. Despite the remarkable potential exhibited by current methods, the inherent data-hungry nature leads to their superior performance being heavily reliant on large-scale annotated data, posing a major challenge in real-world clinical scenarios, as the pixel-wise annotation of medical images is experience-dependent and labor-intensive [Jiao et al., 2023]. To address this problem, researchers have devoted considerable efforts to exploring various data-efficient methods [Feng et al., 2021; Ren et al., 2021; Jiao et al., 2023; Zhang et al., 2023] to achieve higher segmentation performance. As an iterative learning method, Active Learning (AL) can actively select the most valuable or informative samples for annotation during the training process, the purpose of which is to use as little annotated data as possible to achieve optimal model performance [Zhan et al., 2022]. As a result, AL is particularly applicable to medical image segmentation, characterized by high annotation costs and difficulty.

Existing deep AL methods (pool-based) can be categorized into three branches [Zhan et al., 2022]: uncertainty-based, diversity-based, and combined strategies. The core idea of uncertainty-based methods is to query and annotate those samples with high uncertainty. The typical methods [Li and Guo, 2013; Joshi et al., 2009; Brinker, 2003; Wang and Shang, 2014; Kampffmeyer et al., 2016] utilize the predicted posterior probability of the target model to measure uncertainty. However, the overconfidence of deep neural networks often leads to unreliable uncertainty assessments [Zhan et al., 2022]. As shown in Figure 1(a), the uncertainty scores arising from the Maximum Entropy approach [Li and Guo, 2013] fail to reflect the segmentation accuracy of the current
model on unlabeled samples. Some studies [Li and Yin, 2020; Yang et al., 2017; Kim et al., 2023] adopt the bootstrapping strategy to enhance the uncertainty assessment, by leveraging the disagreement of the online committee. AB-UNet [Saidu and Csató, 2021] exploits the Dropout mechanism to emulate a Bayesian network and compute uncertainty according to the Monte Carlo average of multiple forward passes. Despite the performance gain, these methods suffer from significantly increased computational costs, unsuitable for deeper networks and larger datasets. Besides, LPL [Yoo and Kweon, 2019] first proposes a task-agnostic loss prediction strategy to directly predict the loss as the uncertainty of unlabeled samples relative to the target model. However, LPL introduces a joint optimization problem and completely ignores the importance of the predicted posterior probability for uncertainty assessment, resulting in limited performance. Figure 1(b) illustrates that the segmentation loss predicted by the LPL method is highly inconsistent with the actual loss.

Diversity-based approaches aim to query samples that provide varied information for annotation, and most of them use two clustering methods: KMeans [Bodó et al., 2011] and CoreSet [Sener and Savarese, 2018]. KMeans-based methods perform unsupervised clustering on unlabeled samples according to the intermediate features of the target model and then select those samples closest to each centroid while CoreSet-based approaches construct a representative subset as a proxy for the entire dataset. These methods can boost the sample diversity but tend to overlook the informativeness of the acquired samples. Consequently, they are often deemed complementary to uncertainty-based methods, giving rise to a series of combined querying strategies. For instance, Exploration-Exploitation [Yin et al., 2017] builds upon the Maximum Entropy strategy and integrates a determinantal point process to select the most uncertain and diverse samples. BADGE [Ash et al., 2019] proposes a two-stage querying approach. The first stage forms a coarse candidate set based on gradient embedding, while the second stage refines the candidate set through KMeans++ clustering. Other combined strategies [Zhou et al., 2017; Shui et al., 2020] are also designed to uphold both the informativeness and diversity of the selected samples, yet their high complexity entails another engineering cost. Therefore, the primary challenge faced by AL methods in medical image segmentation is: how to more accurately and cost-effectively assess uncertainty while maintaining a balance in the diversity of the selected samples.

To address these challenges, our initial focus is optimizing uncertainty assessment to overcome the shortcomings of existing methods in terms of accuracy and computational efficiency. Inspired by LPL [Yoo and Kweon, 2019], we design a predictive accuracy-driven uncertainty assessment method. The motivation behind it is: if it is possible to predict the loss of a sample point, then why not predict its accuracy relative to the target model? Our preliminary experiments have demonstrated the feasibility of the accuracy prediction, as shown in Figure 1(c). The predicted accuracy of our proposed method exhibits good consistency with the actual accuracy. To this end, we propose a Predictive Accuracy-based Active Learning (PAAL) method for medical image segmentation, first introducing the concept of accuracy prediction. The core idea of PAAL is to use a trained lightweight network to predict the segmentation accuracy of the target model on unlabeled samples, and then guide a diversity-based querying strategy to ensure both uncertainty and diversity of the selected samples.

Specifically, our PAAL mainly consists of an Accuracy Predictor (AP) and a Weighted Polling Strategy (WPS). The AP is a simple neural network that takes the image and the corresponding model predictions as input, aiming to minimize the difference between the predicted and real accuracy. Notably, we design an end-to-end framework to support the simultaneous training of the segmentation model and the attached AP while decoupling their optimization processes. Compared to LPL, our proposed method avoids the joint optimization problem and can leverage the posterior probability to guide the accuracy prediction. Based on this, we design WPS to balance the uncertainty and diversity of the queried samples. After the unsupervised clustering, WPS converts the predicted accuracy of each sample into query weight and cyclically queries the sample with the highest weight in each cluster until iteration ends. Moreover, we propose an Incremental Querying (IQ) mechanism to ensure training stability and facilitate achieving higher performance under a fixed budget. In summary, our contributions mainly include:

- We first propose the concept of Accuracy Predictor (AP) and design a novel active learning method (PAAL) for medical image segmentation. By using the posterior probability as a guide, the attached AP achieves a high consistency between the predicted and actual accuracy, enabling a more accurate measurement of uncertainty.
- We propose a hybrid Weighted Polling Strategy (WPS) to balance the uncertainty and diversity of the acquired samples. Compared to existing methods, our method realizes higher accuracy and more diversified sample distribution, effectively mitigating the issue of imbalanced inter-class annotation.
- Extensive experimental results prove that PAAL outperforms existing methods, achieving accuracy comparable to fully annotated data while reducing annotation costs by approximately 50% to 80%.

2 Related Work

2.1 Active Learning

Active Learning (AL) aims to minimize annotation costs and maximize model performance by selecting the most informative samples for annotation. In this paper, we discuss pool-based active learning methods, which access multiple samples at once. Given an unlabeled sample pool, three main approaches are utilized: uncertainty-based, diversity-based, and combined strategies [Zhan et al., 2022]. Among them, uncertainty-based methods [Li and Guo, 2013; Wang and Shang, 2014; Yuval, 2011; Kampffmeyer et al., 2016] typically use the posterior probability predicted by the target model to define uncertainty. For instance, the Maximum Entropy approach [Li and Guo, 2013] selects those samples with the highest prediction entropy. Due to the overconfidence of deep neural networks, the uncertainty estimation of
such methods is often unreliable. Some studies optimize uncertainty assessment by adopting the bootstrapping strategy [Beluch et al., 2018] or simulating the Bayesian system [Gal et al., 2017; Kendall and Gal, 2017] while introducing higher engineering costs. Diversity-based methods use the intermediate features of the network for unsupervised clustering of samples. These methods can identify the most representative sample points but ignore the informativeness of the selected samples and are thus often considered complementary to uncertainty-based methods. The combined strategies [Yin et al., 2017; Ash et al., 2019] desire to balance the diversity and uncertainty of the acquired samples and have become the major research direction of AL.

2.2 AL for Medical Image Segmentation

Early AL methods are primarily used for image classification. Recently, many researchers have explored the applications of AL methods in medical image segmentation. Due to the differences in network output, most uncertainty-based methods require specific adjustments for segmentation tasks. In contrast, diversity-based methods apply to any task and network since they depend on intermediate features rather than the task-specific output. BioSegment [Rombaut et al., 2022] develops a framework that extends typical AL methods to medical image segmentation tasks. In particular, Li et al. [Li and Yin, 2020] combine the uncertainty assessment based on bootstrapping strategy with similarity representation, proposing a multi-stage combined query strategy. AB-UNet [Saidu and Csató, 2021] adds multiple Dropout layers into the segmentation network to simulate a Bayesian network. It calculates sample uncertainty by obtaining Monte Carlo averages of multiple forward passes. Besides, some studies [Cai et al., 2021; Saidu and Csató, 2019] introduce the concept of super-pixel, which decomposes annotation query from image-level to region-level, attempting to control annotation costs more finely. Other works [Blanch et al., 2017; Zhao et al., 2021] unite the advantages of AL and semi-supervised learning, using high-confidence pseudo-labels to enhance model performance.

3 Methodology

A more accurate uncertainty assessment leads to better performance of AL [Zhan et al., 2022]. As a result, most existing methods for medical image segmentation explore solutions to enhance uncertainty assessment, such as using a bootstrapping strategy [Li and Yin, 2020] or Bayesian network [Saidu and Csató, 2021]. However, these methods exhibit limited performance while significantly increasing computational complexity. The underlying reason is that uncertainty estimation based on the posterior probability can be negatively affected by the overconfidence of the network, as segmentation predictions often contain considerable noise, especially during the early stages of training. LPL [Yoo and Kweon, 2019] suggests utilizing neural networks to model the mapping between the hidden features of images and actual loss. While LPL has shown performance gains in classification tasks, it fails to accurately predict the dense prediction loss based solely on image features and introduces a convergence issue of multi-objective optimization. Inspired by this, we propose a Predictive Accuracy-Based Active Learning (PAAL) method as an alternative learnable uncertainty assessment solution, desiring to overcome the limitations of LPL via simple yet effective designs, as shown in Figure 2.

3.1 Problem Definition

Given an arbitrary medical image segmentation task, let $\mathcal{D}$, $\mathcal{D}_u$, $\mathcal{D}_l$, and $\mathcal{B}$ represent the entire dataset, the unlabeled data pool, the labeled data set, and the specified annotation budget (simply referring to the maximum number of labeled samples), respectively. Following the standard setup of pool-based active learning methods, we have an initial labeled set $\mathcal{D}_l = \{ (x_i, y_i) \}_{i=1}^{M}$ and a large-scale pool of unlabeled samples $\mathcal{D}_u = \{ x_i \}_{i=1}^{N}$, where $M \ll N$, and $x_i$ and $y_i$ represent the $i$-th image and its corresponding true segmentation mask. In the $t$-th iteration of the proposed method, firstly, based on the current segmentation model $\mathcal{M}$, accuracy predictor $\mathcal{P}$, and query strategy $\alpha_{wps} = (\mathcal{D}_u, \mathcal{M}, \mathcal{P})$, a subset $\mathcal{D}_q$ of batch size $b$ is selected from $\mathcal{D}_u$, where $b = \lfloor B/T \rfloor$ and $T$ is the pre-set maximum number of iterations that varies with the annotation budget. Then, we directly query their true labels from the oracle to construct the labeled subset $\mathcal{D}_q^* = \{ (x_i, y_i) \}_{i=1}^{q}$, simulating human annotation. Finally, we update $\mathcal{D}_u = \mathcal{D}_u \setminus \mathcal{D}_q^*$ and $\mathcal{D}_l = \mathcal{D}_l \cup \mathcal{D}_q^*$, and retrain $\mathcal{M}$ and $\mathcal{P}$ using $\mathcal{D}_l$. The iteration process terminates when the budget $B$ is exhausted, and the network converges to a stable state. In this paper, our goal is to maximize the segmentation accuracy of the model using as little labeled data as possible. Since the quality of $\mathcal{D}_l$ is positively related to the performance of $\mathcal{M}$, the key to this study lies in optimizing the query function $\alpha_{wps}$.

3.2 Overview Architecture

Figure 2 illustrates the overall structure and workflow of PAAL. In addition to the basic segmentation network $\mathcal{M}$, PAAL includes two main modules: an Accuracy Predictor $\mathcal{P}$ (AP) and a Weighted Polling Strategy $\alpha_{wps}$ (WPS). The former predicts the segmentation accuracy of the target model for unlabeled samples, while the latter selects a subset $\mathcal{D}_q$ with the most informative and diverse. Without loss of generality, we utilize the U-Net [Ronneberger et al., 2015] with an encoder of ResNet-50 [He et al., 2016] as the base model in the experiments. Unlike LPL, we completely decouple the optimization processes of the AP and the segmentation network, embedding them into an end-to-end unified training framework in a cascaded manner. Besides, our AP utilizes the posterior probability of the segmentation network as prior information, aiming to minimize the discrepancy between the predicted and actual accuracy. More importantly, we design the WPS that utilizes both the predicted accuracy and feature representations of samples to balance the uncertainty and diversity of the acquired samples. The following sections provide detailed explanations of these two core modules.

3.3 Accuracy Predictor

The Accuracy Predictor (AP) of PAAL is fundamentally a regression model in the form of a deep neural network. The primary consideration involves the selection of a suitable accuracy metric as the regression target. Multiple metrics are
available for quantifying segmentation accuracy, such as the Dice Similarity Coefficient (DSC), Hausdorff Distance, and pixel-wise classification accuracy. Considering the stability of convergence, we empirically select a metric with a value range within $[0, 1]$ to represent “Predictive Accuracy”. We use DSC as the regression target in the experiments, although it can be replaced with any normalized metric. To reduce computation, AP is a simple variant of ResNet-18, which adds a Sigmoid layer behind the linear layer of the classification head. During training, we concatenate the input image $O_0 \in \mathbb{R}^{B \times C' \times H \times W}$ and its corresponding segmentation prediction probability $O_1 \in \mathbb{R}^{B \times C' \times H \times W}$ along the channel dimension and deliver it to AP, where $B$, $C'$, $H$, $W$, and $C$ denote the batch size, number of image channels, height and width of the image, and the number of segmentation categories, respectively. The optimization process utilizes Mean Squared Error (MSE) loss $L_{mse}$ to minimize the difference between the predicted $O_2 \in \mathbb{R}^{B \times C}$ and actual accuracy. Notably, during the early stages of joint training, we set a brief silent period (5 epochs) for AP, meaning that only the segmentation network undergoes training during this period to alleviate the impact of early segmentation noise. Compared to the loss prediction module of LPL, the optimization process of the proposed AP is independent of the segmentation network, thereby bypassing the multi-objective optimization problem. Further, utilizing the posterior probability of the segmentation network, rather than solely relying on image features, as prior information helps to enhance convergence. Experimental results showcase a high consistency between predicted and actual accuracy, affirming the effectiveness of the proposed accuracy prediction approach.

### 3.4 Weighted Polling Strategy

From the perspective of data efficiency, the pixel-level annotation of medical images faces two challenges: sample redundancy and imbalanced inter-class annotation. Taking 3D images as an example, a CT or MR dataset typically contains numerous highly similar slices due to the similarity of anatomical structures, leading to redundant annotations and reduced data efficiency. Moreover, there are significant volume differences between tissues or organs, and small-volume targets only appear in a few slices, causing an imbalanced annotation distribution that may hamper the segmentation accuracy of the model for minority classes. Although AP can identify the samples with high uncertainty, it fails to ensure their diversity. To address these issues, we propose a hybrid Weighted Polling Strategy (WPS) to balance the informativeness and diversity of the selected samples.

As illustrated in Algorithm 1, in the query process of active learning, WPS first transforms the predicted accuracy of unlabeled samples into query weights $W$. The query weight $w_i$ for the $i$-th sample is negatively correlated with the overall predicted accuracy. The computation is detailed as follows:

$$W = \{w_i\}_{i=1}^{N}, w_i = \frac{1}{C} \sum_{j=1}^{C} -\log(p_j^i),$$

where $p_j^i$ denotes the predicted accuracy for the $j$-th segmentation class of the $i$-th sample. We employ the logarithmic mean to amplify attention to minority classes. Then, based on the hidden features of the segmentation model, a naive KMeans algorithm is utilized for unsupervised clustering of the unlabeled samples, yielding $K$ clusters $\{\Omega_j\}_{j=1}^{K}$. To reduce computational complexity, we set a small $K = \lceil \log_2(b+4) + 1 \rceil$ and use an adaptive global average pooling layer to compress the original representations. Finally, we alternately query the sample with the highest weight in each cluster until the current iteration concludes. Compared to existing diversity-based methods, WPS exhibits lower computational complexity, suitable for deeper networks and larger datasets, and considers both the uncertainty and distribution of the selected samples to alleviate the issue of imbalanced inter-class annotation. Besides, we propose an Incremental Querying (IQ) mechanism that differs from querying based
Algorithm 1 The Proposed PAAL Process

Input: Unlabeled dataset \( D_u \), Initial labeled dataset \( D_l \), Segmentation network \( M \), Accuracy predictor \( P \), Oracle
Parameter: Maximum iterations \( T \), Number of clusters \( K \)
Output: Final \( D_f \), \( M \) and \( P \)

1: \( t \leftarrow 1 \), \( IQ \leftarrow 0 \)
2: while not reach the budget and stable convergence do
3: \( M, P \leftarrow D_l \)
4: if \( t \leq T \) and \( IQ \geq 10 \) then
5: \( O_1, F \leftarrow (M, D_u) \)
6: \( O_2 \leftarrow (P, O_1, D_u) \)
7: \( W \leftarrow O_2 \)
8: \( \{D^t_q\}_{q=1}^K \leftarrow (F, K) \)
9: \( D^t_q \leftarrow (\{O_i\}_{i=1}^K, W) \)
10: \( D^t_q \leftarrow \text{Oracle}(D^t_q) \)
11: \( D_1 \leftarrow D_l \cup \bigcup_{q=1}^K \{D^t_q \} \)
12: \( D_u \leftarrow D_u \setminus \{D^t_q \} \)
13: \( t \leftarrow t + 1 \), \( IQ \leftarrow 0 \)
14: end if
15: if not \( M \uparrow \) then
16: \( IQ \leftarrow IQ + 1 \)
17: else
18: \( IQ \leftarrow 0 \)
19: end if
20: end while
21: return \( D_f \), \( M \) and \( P \)

4 Experiments and Results

4.1 Datasets

As shown in Table 1, datasets used in our experiments include (1) Brain Tumour [Antonelli et al., 2022]: a multi-modal Magnetic Resonance (MR) dataset provided by the Medical Segmentation Decathlon (MSD), comprising 484 annotated samples with segmentation targets of brain Edema, Enhanced (ET) and non-Enhanced tumors (nET); (2) SegTHOR [Lambert et al., 2020]: a chest Computed Tomography (CT) dataset containing only 40 scans, annotated for 4 organs; (3) ACDC [Bernard et al., 2018]: a commonly used cardiac MR dataset composed of scan images from 100 patients, annotated for the Left Ventricle (LV), Right Ventricle (RV), and Myocardium (Myo). More importantly, to explore the application potential of the proposed method, we constructed (4) Liver OAR: a clinical organ-at-risk (OAR) segmentation dataset for liver cancer, annotated for 8 abdominal organs, collected by the Radiotherapy Department of the First Affiliated Hospital of University of Science and Technology of China, where all CT images were annotated and verified by two experienced physicists and have been used in radiotherapy planning. We set different initial annotation ratios for different datasets due to the varying slice scales.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Modality</th>
<th>Samples (slices)</th>
<th>Init.R.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brain Tumour</td>
<td>multi-MR</td>
<td>484 (66,512)</td>
<td>0.5%</td>
</tr>
<tr>
<td>SegTHOR</td>
<td>CT</td>
<td>40 (7,420)</td>
<td>5.0%</td>
</tr>
<tr>
<td>ACDC</td>
<td>MR</td>
<td>100 (1,902)</td>
<td>5.0%</td>
</tr>
<tr>
<td>Liver OAR</td>
<td>CT</td>
<td>49 (3,725)</td>
<td>5.0%</td>
</tr>
</tbody>
</table>

Table 1: Datasets used in the experiments. * denotes the private dataset from clinical, and Init.R. refers to the initial annotation ratio.

Specifically, we list the annotation ratio for different categories on different datasets as follows. For Brain Tumour, (Edema, nET, ET) = (50%, 35%, 31%); for SegTHOR, (Esophagus, Heart, Trachea, Aorta) = (53%, 21%, 27%, 51%); for ACDC, (RV, Myo, LV) = (82%, 95%, 95%); and for Liver OAR, (Spinal-Cord, Small-Intestine, Kidney-L, Kidney-R, Liver, Heart, Lung-L, Lung-R) = (80%, 44%, 22%, 20%, 30%, 17%, 34%, 34%). We can see that except for the ACDC dataset, the other datasets suffer from different degrees of inter-class annotation imbalance. During training, each dataset was split into training and validation sets at a ratio of 8:2 for five-fold cross-validation. All reported DSC results in subsequent sections are the average and standard deviation of the five-fold.

4.2 Implementation Details

PAAL and all baselines are implemented using PyTorch and integrated into a unified training framework. In particular, we select representative methods as baselines, including uncertainty-based, diversity-based, and combined methods, all of which have open-source implementations. All models are trained from scratch on 8 NVIDIA A800 GPUs, with the same loss function, e.g., the combined loss [Shi et al., 2023] of Dice and Cross-Entropy for segmentation model and MSE loss for AP. We set 3 maximum querying ratios for different datasets according to varying slice scales: \{5%, 10%, 20%\} for the Brain Tumour dataset and \{10%, 20%, 50%\} for the other datasets. Unlike the proposed IQ mechanism, the query interval for comparison methods is set to 5 epochs. The maximum iteration for each dataset is related to the maximum querying ratio. Specifically, Brain Tumour dataset has maximum iterations of \{10, 15, 15\}, while the other datasets have the same \{5, 15, 20\}. For the Brain Tumour dataset, the slice resolution is resized to \( 4 \times 256 \times 256 \), while for the other datasets, it is \( 1 \times 512 \times 512 \). We employ AdamW optimizer [Loshchilov and Hutter, 2018] with an initial learning rate of 1e-3, a batch size of 64, and use the cosine annealing strategy [Loshchilov and Hutter, 2016] to control the learning rate, with a weight decay of 1e-4, warm-up epochs of 10, and the minimum learning rate of 1e-6. Each model is evaluated on the validation set at the end of every epoch. To alleviate overfitting, we adopt an early stopping strategy with a tolerance of 40 epochs to search for the best model within 400 epochs and apply data augmentation, including random distortion, rotation, flip, and noise.

4.3 Overall Performance

Results on Open-Source Datasets. In Table 2, we report the results of the proposed PAAL on open-source single-
modal datasets ACDC [Bernard et al., 2018] and SegTHOR [Lambert et al., 2020], as well as the multi-modal dataset Brain Tumour [Antonelli et al., 2022], compared with state-of-the-art methods. PAAL (only AP) indicates the removal of the WPS module, selecting samples solely based on query weights, similar to other uncertainty-based methods. We can see that PAAL significantly outperforms all previous methods, achieving the highest DSC across different datasets with varying annotation ratios. In particular, at the lowest annotation budget, our proposed method surpasses typical Maximum Entropy [Li and Guo, 2013], KMeans [Rombaut et al., 2022], and LPL [Yoo and Kweon, 2019] methods by 3.8%, 2.2%, and 15.9% on ACDC, 9.2%, 3.1%, and 8.9% on SegTHOR, and 4.9%, 0.9%, and 20.7% on Brain Tumour, respectively. These results demonstrate the superior data efficiency of PAAL under a limited budget. Notably, the performance differences among different methods diminish as the annotation ratio increases. PAAL achieves segmentation accuracy comparable to fully annotated data at annotation ratios of 50% and 20%, respectively.

An intriguing observation is that most uncertainty-based methods relying on the posterior probability perform even worse than random sampling, implying their limited applicability to segmentation tasks. We hypothesize that this phenomenon stems from the potential noise introduced by network overconfidence, making uncertainty assessment prone to failure in dense prediction tasks. The experimental results provide supporting evidence. For example, AB-UNet [Saidu and Csató, 2021] uses a Bayesian network to enhance uncertainty assessment and achieves the performance gain, while CEAL [Blanch et al., 2017] using pseudo-labels performs worse in most cases, indicating the low reliability of the network predictions. Although diversity-based methods [Rombaut et al., 2022; Zhan et al., 2022] can maintain relatively satisfactory performance, they are constrained by network depth and data scale. Moreover, LPL performs poorly due to convergence issues arising from joint optimization, especially on the complicated Brain Tumour Datasets. In contrast, our proposed PAAL outperforms all comparison methods even using AP alone, demonstrating its effectiveness.

### Results on Private Clinical Dataset

Table 3 shows the results on the small-scale private dataset, Liver OAR. Similarly, PAAL achieves the highest DSC at 10%, 20%, and 50% annotation ratios, reaching 89.7%, 90.8%, and 91.9%, respectively. Notably, at an annotation budget of 20%, except for our proposed method, CoreSet, and AB-UNet, all other comparative methods perform worse than random sampling. Given the computational complexity, CoreSet is unsuitable for deeper networks and larger datasets, and AB-UNet also requires additional computations to simulate the Bayesian network. In contrast, PAAL achieves higher segmentation performance with lower computational overhead, especially when using only AP. These results further validate the effectiveness of PAAL in reducing annotation costs, showcasing significant potential in practical applications.

### Time Efficiency Analysis

The average query time in Table 3 demonstrates the time efficiency of different methods. The query time of PAAL mainly consists of the inference time and the query time of PAAL mainly consists of the inference time, which is significantly lower than that of the diversity-based methods. After WPS removal,
Table 4: Ablation study on ACDC dataset, w/o denotes without.

<table>
<thead>
<tr>
<th>Method</th>
<th>ACDC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10%</td>
</tr>
<tr>
<td>Random</td>
<td>85.3±2.9</td>
</tr>
<tr>
<td>w/o WPS and IQ</td>
<td>85.3±2.5</td>
</tr>
<tr>
<td>w/o IQ</td>
<td>86.4±2.6</td>
</tr>
<tr>
<td>w/o WPS</td>
<td>86.3±2.5</td>
</tr>
<tr>
<td>w/o AP</td>
<td>84.6±4.3</td>
</tr>
<tr>
<td>PAAL</td>
<td>86.8±2.2</td>
</tr>
<tr>
<td>Full data</td>
<td>91.6±1.4</td>
</tr>
</tbody>
</table>

Figure 3: The annotation distribution of different methods.

Figure 4: The performance curves under different labeled ratios.

maximum annotation budget (20% for Brain Tumour, 50% for the others). Notably, we introduce a background category denoted as B.g., to represent those images without any segmentation target. It can be observed that, compared to the original distribution represented by Random, different methods exhibit significant differences. For the ACDC and Brain Tumour datasets, PAAL achieves better performance by significantly increasing the annotation ratio of minority classes or reducing the ratio of majority classes. For example, the annotated slice ratio of “Edema” and “ET” is reduced from 1.62 to 1.26. As for the other two datasets, the annotation distribution of PAAL is generally consistent with the original distribution, while LPL and Maximum Entropy show significant differences, leading to poor performance. These results demonstrate that PAAL can adaptively adjust based on the data distribution of the specific task, helping to alleviate the problem of imbalanced annotation of multiple categories.

Furthermore, we present the performance curves of different query strategies as the annotation ratio iteratively increases under the maximum annotation budget. As shown in Figure 4, unlike the drastic fluctuations of the existing methods, the DSC rising curve of the proposed PAAL is remarkably smooth. The essential reason is that the proposed IQ mechanism ensures training stability by triggering queries only after achieving the best performance on current data. Besides, PAAL consistently maintains the best segmentation performance under different annotation ratios, further proving its effectiveness and superiority.

5 Conclusion

In this paper, we proposed a Predictive Accuracy-based Active Learning (PAAL) approach for medical image segmentation. Specifically, we employed a lightweight Accuracy Predictor (AP) to directly predict the segmentation accuracy of

4.4 Ablation Study
To reveal the effect of different modules on performance improvement, we conduct ablation studies on AP, WPS, and IQ modules on the ACDC dataset, and the results are shown in Table 4. After removing each module separately, there is a varying degree of performance decline, indicating the essential importance of all modules for PAAL. It is worth noting that w/o AP means that PAAL degenerates into the KMeans method, where the most significant performance degradation occurs. In low-budget scenarios, the IQ leads to a notable improvement in overall performance, demonstrating the effectiveness of IQ in enhancing network convergence. Furthermore, we can observe that both AP and WPS have a significant impact on overall performance, further highlighting the superiority of our designs.

4.5 Quantitative Analysis
As shown in Figure 3, we compare the annotation distribution of samples selected by different query strategies under the time efficiency of our proposed method is better than that of all comparison methods due to the lightweight structure of AP. Overall, PAAL achieves a good trade-off in terms of accuracy and time efficiency.
unlabeled samples related to the target model, and designed a hybrid Weighted Polling Strategy (WPS) to balance uncertainty and diversity. Extensive experimental results demonstrated the superiority of PAAL over existing methods. The low complexity and high data efficiency of PAAL indicated significant potential for clinical applications. In the future, we will explore more optimization methods such as semi-supervised learning to further enhance the performance.

Acknowledgments

This work is financially supported by the National Key Research and Development Program of China (Grants No. 2016YFB1000403), the Fundamental Research Funds for the Central Universities (Grant No. YD2150002001), and the National Natural Science Foundation of China (Grant No. 60970023).

Contribution Statement

Jun Shi and Shulan Ruan have equal contributions to this paper. Hong An is the corresponding author.

References


