TECHNICAL CONTRIBUTION

Partial Image Active Annotation (PIAA): An Efficient Active Learning Technique Using Edge Information in Limited Data Scenarios

Md Abdul Kadir[1,2] · Hasan Md Tusfiqur Alam[1] · Devansh Srivastav[1] · Hans-Jürgen Profitlich[1] · Daniel Sonntag[1,2]

Abstract

Active learning (AL) algorithms are increasingly being used to train models with limited data for annotation tasks. However, the selection of data for AL is a complex issue due to the restricted information on unseen data. To tackle this problem, a technique we refer to as Partial Image Active Annotation (PIAA) employs the edge information of unseen images as prior knowledge to gauge uncertainty. This uncertainty is determined by examining the divergence and entropy in model predictions across edges. The resulting measure is then applied to choose superpixels from input images for active annotation. We demonstrate the effectiveness of PIAA in multi-class Optical Coherence Tomography (OCT) segmentation tasks, attaining a Dice score comparable to state-of-the-art OCT segmentation algorithms trained with extensive annotated data. Concurrently, we successfully reduce annotation label costs to 12%, 2.3%, and 3%, respectively, across three publicly accessible datasets (Duke, AROI, and UMN).

Keywords Active learning · Deep learning · Segmentation · Medical image · OCT · AL

1 Introduction

In recent years, Deep Learning (DL) based methods have achieved considerable success in medical image segmentation [1]. However, their progress has often been constrained as they require large datasets. AL has the potential to significantly enhance the efficiency of any intelligent diagnostic system such as [2] by mitigating the need for extensive annotation efforts, as evidenced in previous studies [3, 4]. For example, ophthalmologists use the segmentation of ocular OCT images for diagnosis, and treatment of eye diseases such as Diabetic Retinopathy (DR) and Diabetic Macular Edema (DME) [5, 6]. Labeling medical image data for AL is a time-consuming and expensive process as domain experts are required to annotate them manually. In this study, we primarily rely on data derived from OCT. This technology has gained significant popularity in the field of ophthalmology imaging due to its effectiveness. OCT employs the use

of light waves, enabling it to generate high-definition, cross-sectional visuals of the internal structures of the eyes.

In ophthalmology, OCT is used to diagnose and monitor conditions such as macular degeneration, diabetic retinopathy, and glaucoma. OCT images can provide detailed information about the thickness and integrity of retinal layers, the presence of fluid or swelling, and the size and shape of optic nerve structures. To observe the development and changes in retinal layers, the presence of fluid or swelling, and the size and shape of optic nerve structures during the treatment phase, doctors annotate the images. This allows them to easily track how the retinal layers change over the course of treatment.

AL can serve as a beneficial tool in the realm of medical image segmentation. It has the potential to alleviate the extensive effort required for annotation by leveraging the model to obtain annotations for image regions where the model exhibits high confidence. Conversely, in instances where the model demonstrates lower confidence, experts can contribute by providing more ground truth data [3]. In practice, expert annotation of large-scale medical image databases is highly laborious, resource-intensive, and often infeasible.

In response to this challenge, we propose PIAA, a region-based AL technique. This technique, which trains over time

✉ Md Abdul Kadir
abdul.kadir@dfki.de

1 German Research Center for Artificial Intelligence, Kaiserslautern, Germany

2 University of Oldenburg, Oldenburg, Germany

using a minimal amount of annotated regions, aids ophthalmologists by generating OCT segments. The PIAA framework capitalizes on the prediction uncertainty across the boundaries of the semantic regions of input images, informing the end user about the segmentation areas it is confident about and of those it is not. The end user only accepts segmentation output from the confident areas while providing feedback to the model on less confident areas. The model learns based on this feedback, and as a result, its performance improves over time. Edge information is one of the image's most salient features, and it can boost segmentation accuracy when integrated into neural model training [7]. We formulate a novel acquisition function that leverages the variance of the predicted score across the gradient surface of the input to measure uncertainty.

Empirical results show that PIAA outperforms other state-of-the-art AL methods in three OCT image datasets. This paper is an extended version of EdgeAL [8], offering additional experimental results, illustrations, and practical use cases to provide a more comprehensive elaboration of the methodology, experiments, and results of the algorithm. The extended content is presented to enhance the understanding of the algorithm's capabilities and use cases.

2 Related work

Active learning is applied to a variety of tasks, including natural language processing, computer vision, and reinforcement learning, and it is expected to play a major role in the development of interactive machine learning methods. It is a cost-effective method that selects the most informative samples for annotation to improve model performance based on uncertainty [9], data distribution [10], expected model change [11], and other criteria [12]. A simpler way to define uncertainty is to use the posterior probability of the predictions, e.g., to select an instance with the least confident posterior probability [9, 13] or the margin between posterior probabilities for different predicted class [14, 15]. Some methods [16, 17] use the entropy of class posterior as an uncertainty measure. These methods are often used in conjunction with sampling-based strategies to estimate model uncertainty which is based on the inconsistency of predictions [3, 12, 18].

In the context of active learning, when dealing with a pool of unlabeled data, there are primarily three major strategies that can be employed to select the next batch of data that needs to be labeled. These strategies include uncertainty-based approaches, distribution-based approaches, and methods based on expected model change [19].

In the uncertainty-based approach, the learning algorithm seeks out samples that carry the highest degree of prediction uncertainty. This can be gauged by measuring the posterior probability of a predicted class [13, 20]. The fundamental belief underpinning this approach is that these samples, once they have been labeled, hold the potential to provide the most critical information that can enhance the learning capacity of the model.

The distribution-based approach in active learning is centered on the selection of data points that embody the entire distribution of an unlabeled data pool. The underlying premise of this approach is that learning from a representative subset of the data can yield results that are as competitive as learning from the entire data pool. This approach can be implemented in several ways. For instance, Nguyen and Smeulders [21] employ a clustering algorithm to partition the data pool, thereby facilitating the identification of representative data points. Alternatively, Yang et al. [22], Guo [23] and Elhamifar et al. [24] formulate the selection of a representative subset as a discrete optimization problem, thus ensuring the selection of the most informative data points. Another approach, as proposed by [25, 26], involves evaluating the proximity of a data point to its surrounding data points, thereby selecting data points that can effectively propagate knowledge across the dataset. These methods, therefore, ensure the selection of the most representative and informative data points for model training, thereby optimizing the active learning process.

The technique of the expected model change is a more advanced and decision-theoretic strategy for model enhancement. This technique utilizes the existing model to predict the expected length of the gradient [27], anticipated future errors [28], or predicted changes in output [29] for all potential labels. These strategies, initially designed for use with smaller models and datasets, can be assessed for their efficiency when applied to larger deep networks [30, 31] and extensive datasets [32].

The approach based on uncertainty [13, 33] has shown robust results for classification tasks. However, it requires a task-specific design for other tasks as it leverages network outputs. In a more general approach, Gal et al. [34] achieves uncertainty estimates via multiple forward passes using Monte Carlo Dropout. Although this method has been validated with small-scale classification tasks, it is computationally demanding for recent large-scale learning due to the need for dense dropout layers, which significantly slow down the convergence rate. Beluch et al. [35] proposes an ensemble method that consists of 5 deep networks to measure uncertainty through disagreement. While it has demonstrated cutting-edge classification performance, it is not efficient in terms of memory and computation for large-scale challenges.

Many AL methods are adopted for segmentation tasks [14, 36, 37]. Gorriz et al. [36] propose an AL framework Melanoma segmentation by extending Cost-Effective Active Learning (CEAL) [38] algorithm where complimentary samples of both high and low confidence are selected for
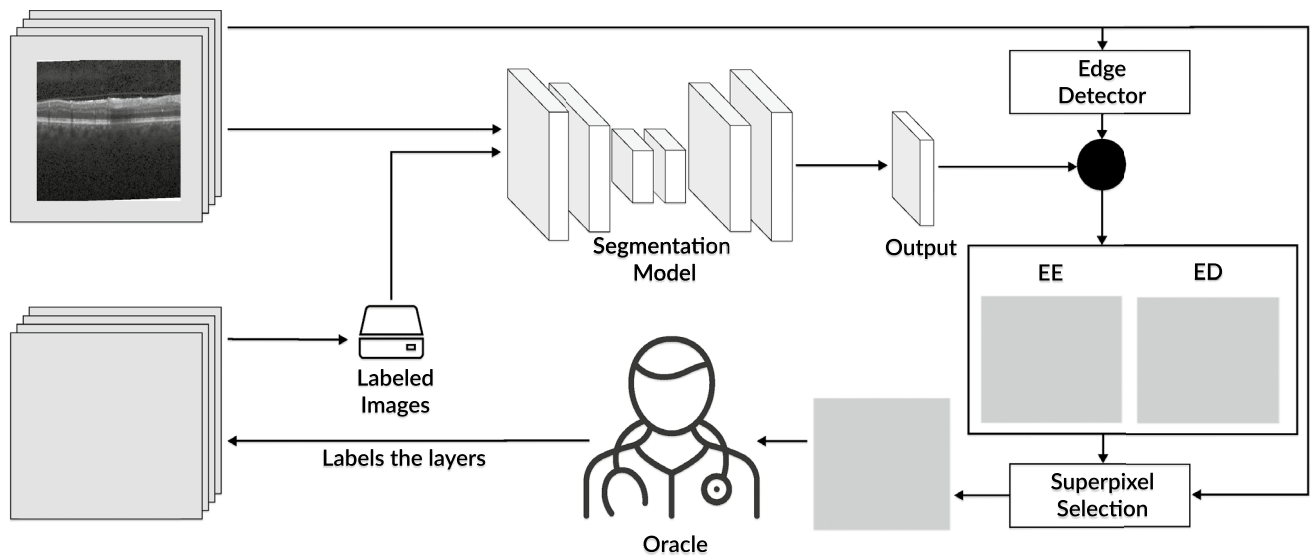
Fig. 1 Workflow of our algorithm. Given OCT images, by using Monte-Carlo simulation our algorithm first computes the edge entropy (EE) and edge divergence (ED) maps of the outputs of the segmentation model. Based on these maps, it calculates the overlaps between superpixels and recommends the annotation regions. The algorithm then utilizes the recommended region to guide the annotation process, with the annotation being stored for future trainings

annotation. Mackowiak et al. [37] use a region-based selection approach and estimate model uncertainty using MC dropout to reduce human-annotation cost. Nath et al. [14] propose an ensemble-based method where multiple AL frameworks are jointly optimized, and a *query-by-committee* approach is adopted for sample selection. These methods overlook the incorporation of prior information, such as the structure of the image, edge details, and morphological data, in their uncertainty estimation process. Authors in [39] propose an AL framework for multi-view datasets [40] segmentation task where model uncertainty is estimated based on *Kullback–Leibler divergence* (*KL-divergence*) of posterior probability distributions for a disjoint subset of prior features such as depth, and camera position.

However, this viewpoint information is not always readily available in medical imaging, and even when it is, it may not make a significant difference. This is largely due to the static positioning of most medical imaging devices, which limits the variability and potential impact of different viewpoints. We leverage edge information as a prior for AL sampling based on previous studies where edge information has improved the performance of segmentation tasks [7]. To the best of our knowledge, while numerous classical computer vision research studies have demonstrated that edge detection methods can be utilized for segmentation [41], there has not yet been any exploration of using image edges as an *a priori* in active learning.

Moreover, there is not sufficient research other than [42] related to Active Learning for OCT segmentation. The proposed approach in [42] requires foundation models [31] to be pre-trained on large-scale datasets in similar domains, which can be infeasible to collect due to data privacy. On the other hand, our method requires only few samples (~2% of the usual subset) for initial training, overcoming the limitation of the need for a large dataset.

3 Methodology

Figure 1 illustrates our active learning method which comprises four key phases. First, we initiate network training using a subset of labeled images, typically a small fraction of the entire dataset (e.g., 2%). Next, we calculate uncertainty metrics for both individual input instances and specific input regions, and using this information we make choices regarding which superpixels to annotate, and we acquire annotations through a simulated oracle.

3.1 Segmentation Network

At first, we train our OCT semantic segmentation model by selecting a small, random subset of labeled data D_s, which is used as the seed set. The remainder of the labeled data is used to simulate an Oracle. For our primary architecture, we use Y-net-gen-ffc (YN*), and we choose not to initialize it with pre-trained weights due to its documented superior performance [5].

Y-Net is composed of two distinct encoder branches: the *spatial encoder* for convolutional blocks and the *spectral encoder* to integrate fast Fourier convolutional (FFC) blocks
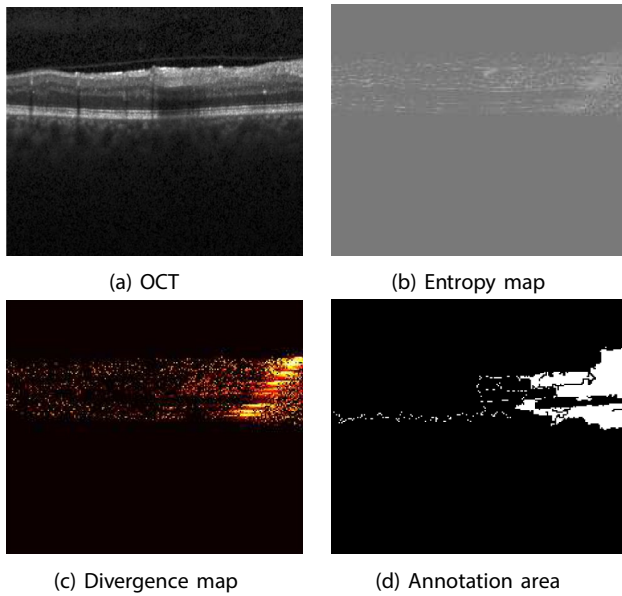
(a) OCT

(b) Entropy map

(c) Divergence map

(d) Annotation area

Fig. 2 Illustrations of an instance of **a** an OCT slice along with its associated **b** *edge entropy* map, **c** *edge divergence* map, and **d** designated query region by our PIAA method. These visual representations highlight a notable observation: the right side of the OCT slice exhibits its reduced clarity in retinal layer separation lines. This diminished clarity could potentially account for the model's heightened uncertainty within that specific region

[43]. The decoding process is handled by a single decoder architecture, which takes the spatial and spectral features extracted by the encoder networks as input and generates the segmentation map. Similar to the architectural design of U-Net [44], Y-Net follows an autoencoder-based structure, incorporating skip connections that link spatial encoder blocks with decoder blocks. The function of the spectral encoder is to identify and handle global features originating from the frequency domain, which could be overlooked when depending exclusively on spatial convolutions.

Moreover, we also train DeepLabv3 and U-net models with ResNet and MobileNetv2 as encoder backbone for ablation experiments. For these models, we conduct experiments with ImageNet [32] and Kaiming [45] weight initialization.

3.2 Computing Prediction Uncertainity

PIAA aims to enhance the model's performance by actively querying uncertain regions within unlabeled data D_u following its training on an initial dataset D_s. These uncertain regions are believed to be particularly valuable for further training. To achieve this goal, we introduce a novel edge-based uncertainty measurement strategy. This approach involves the computation of two key metrics: the *edge entropy* score and the *edge divergence* score. They are utilized to assess the prediction ambiguity associated with the edges between layers in the OCT images. Figure 2 provides visual examples of input

OCT data along with the measured *edge entropy* and *edge KL-divergence* corresponding to the input.

3.2.1 Edge Entropy Score

Analyzing the edges of raw OCT (Optical Coherence Tomography) inputs can provide crucial insights into image features and texture. While these edges may appear noisy at first glance, they serve as a concise representation of all the alterations present in an image. The Sobel operator, as detailed in [7], is a suitable tool for detecting edges in the input image. Let's define the normalized absolute value of edges in an image I_i of size (M, N) as S_i. In this context, $|\nabla I_i|$ represents the absolute gradient, and S_i is calculated using the following equation:

$$S_i^{m,n} = \frac{|\nabla I_i^{m,n}| - \min(|\nabla I_i|)}{\max(|\nabla I_i|) - \min(|\nabla I_i|)} \quad \text{for } m, n \in M, N$$

$\min(|\nabla I_i|)$ and $\max(|\nabla I_i|)$ represent the minimum and maximum values, respectively, within the absolute gradient matrix $|\nabla I_i|$.

Additionally, to assess the probability that each pixel in an image belongs to a specific class denoted as c, we rely on the network's output, represented as $P_i^{(m,n)}(c)$. To introduce uncertainty into our observations, we adopt a Monte Carlo (MC) dropout simulation method as outlined in [34]. This involves averaging predictions over $|D|$ occurrences. Consequently, an MC probability distribution indicates the likelihood of a pixel at position (m, n) in the image I_i belonging to class c, where C represents the set of segmentation classes. During the neural network evaluation phase, we run MC dropouts $|D|$ times and measure $P_i^{(m,n)}(c)$ using Eq. (1):

$$P_i^{(m,n)}(c) = \frac{1}{|D|} \sum_{d=1}^{D} P_{i,d}^{(m,n)}(c) \tag{1}$$

Following [46], we apply contextual calibration on $P_i^{(m,n)}(c)$ by S_i to prioritize significant input surface variations. Now, S_i is linked with a probability distribution, with $\phi_i^{(m,n)}(c)$ having information about the edges of input. This formulation makes our implementation unique from other active learning methods in image segmentation.

$$\phi_i^{m,n}(c) = \frac{e^{P_i^{(m,n)}(c) \cdot S_i(m,n)}}{\sum_{k \in C} e^{P_i^{(m,n)}(k) \cdot S_i^{(m,n)}}} \tag{2}$$

We name $\phi_i^{m,n}(c)$ as contextual probability and define our *edge entropy* by following the entropy formula of [17].

$$EE_i^{m,n} = -\sum_{c \in C} \phi_i^{m,n}(c) \log(\phi_i^{m,n}(c)) \tag{3}$$

3.2.2 Edge Divergence Score

In regions with pronounced edges or gradients, the *edge entropy* metric signifies the extent of inconsistency in the network's predictions for each individual pixel in the input. Nevertheless, it is imperative to quantify the extent of this uncertainty. To achieve this, we employ the concept of *KL-divergence* to quantify the dissimilarity between $P_i^{(m,n)}$ and $\phi_i^{(m,n)}$ for a specific pixel located at coordinates (m, n) within an input image. This approach is based on the concept of self-knowledge distillation within the context of I_i [47]. The *edge divergence* score, denoted as $ED_i^{(m,n)}$, can be formally defined using Eqs. 1 and 2 as follows:

$$ED_i^{(m,n)} = D_{KL}\big(P_i^{(m,n)}||\phi_i^{(m,n)}\big)$$

Here $D_{KL}\big(P_i^{(m,n)}||\phi_i^{(m,n)}\big)$ quantifies the distinction between the model's predictive probability and the contextual probability for pixels belonging to the edges of the input (Fig. 2).

3.3 Superpixel Selection

Clinical images often have a sparse representation, and the critical or relevant information is localized in a small portion of the image. This characteristic can be particularly advantageous for active learning-based annotations allowing experts to concentrate on the most informative areas [37]. We use a traditional segmentation technique, SEEDS [48], to leverage the local structure from images for finding superpixels. Annotating superpixels and regions for active learning may be more beneficial to the user than annotating the entire picture [37].

We calculate the mean *edge entropy* EE_i^r and mean *edge divergence* ED_i^d for a certain area r within a superpixel. These can be expressed as follows:

$$EE_i^r = \frac{1}{|r|} \sum_{(m,n)\in r} EE_i^{(m,n)} \tag{4}$$

$$ED_i^r = \frac{1}{|r|} \sum_{(m,n)\in r} ED_i^{(m,n)} \tag{5}$$

Here, $|r|$ represents the number of pixels within the superpixel region.

We use regional entropy to identify the optimal superpixel for our selection strategy, selecting the one with the highest value based on [39].

$$(i, r) = \underset{(j,s)}{\arg\max} \quad EE_j^s \tag{6}$$

Following [39], we identify a subset of superpixels in the dataset with a 50% overlap, forming a set, R. We choose the

superpixels with the largest *edge divergence* to determine the ultimate query (sample) for annotation:

$$(p, q) = \underset{(j,s)\in R}{\arg\max} \big\{ ED_j^s \quad | \quad (j,s) \cap (i,r); (i,r) \in D_u) \big\} \tag{7}$$

After each selection, we remove the chosen superpixels from set R. This selection process continues until we have selected a total of K superpixels from set R.

In the edge case, when gradients are absent in the image, *edge divergence* ($ED_i^{(m,n)}$) is assigned constant for every pixel, as the *contextual probability* ($\phi_i^{(m,n)}$) becomes the same as the original probability ($P_i^{(m,n)}$). This results in *edge divergence* being zero for every pixel. However, *edge entropy* ($EE_i^{(m,n)}$) is calculated based on calibrated probability ($\phi_i^{(m,n)}$), which is basically $P_i^{(m,n)}$.

During the process of query selection, we initially arrange superpixels based on edge entropy (EE_j^s). From this set of superpixels, we select the one with the highest edge divergence (ED_j^s). In cases where all divergence values are the same (zero), the selection of queries is determined randomly based on divergence values. Nevertheless, since the initial criterion for superpixel selection was the order of EE_j^s, the entropy of a superpixel becomes the decisive factor in query selection criteria. If edges are present in the image, the query criteria are determined by *edge divergence*. On the other hand, in the absence of edges, the criteria shift to the entropy of the model MC output, $P_i^{(m,n)}$.

3.4 Simulated Labeling (Oracle)

A simulated annotator is used to label the ground truth for our active learning system. This virtual annotator is known as the oracle, and it has access to all of the ground truth label information, as illustrated in Fig. 1. Upon obtaining the selected superpixel maps described in Sect. 3.3, we acquire the corresponding ground truth information for those regions from the oracle. At each active iteration labeled data set expands with the annotated data and the unlabeled data set shrinks. In each active learning iteration, the model is freshly trained on the updated dataset.

4 Experiments and Results

In this section, we give a comprehensive overview of the datasets and architectures utilized in our experiments. Then, we present our extensive experimental results and compare them with results from other state-of-the-art methods to illustrate the effectiveness of our approach. We compare our AL method with nine other well-established active learning strategies: softmax margin (MAR) [15], softmax confidence (CONF) [38], MC dropout entropy (MCDR) [34], softmax

Table 1 Results from a fivefold cross-validation (mean Dice scores ± standard deviation) for the PIAA and other active learning techniques method on the Duke dataset for the YN* segmentation Model

GT(%)	2%	12%	22%	33%	43%	100%
RMCDR	0.40 ± 0.05	0.44 ± 0.04	0.63 ± 0.05	0.58 ± 0.07	0.70 ± 0.03	0.82 ± 0.03
CEAL	0.40 ± 0.05	0.54 ± 0.04	0.54 ± 0.04	0.55 ± 0.06	0.79 ± 0.03	0.82 ± 0.03
CORESET	0.38 ± 0.04	0.44 ± 0.05	0.62 ± 0.04	0.57 ± 0.04	0.69 ± 0.03	0.82 ± 0.03
PIAA	**0.40 ± 0.05**	**0.82 ± 0.03**	**0.83 ± 0.03**	**0.81 ± 0.04**	**0.83 ± 0.02**	0.82 ± 0.02
MAR	0.40 ± 0.09	0.44 ± 0.03	0.58 ± 0.04	0.67 ± 0.03	0.70 ± 0.04	0.83 ± 0.02
MAXRPR	0.41 ± 0.04	0.54 ± 0.09	0.67 ± 0.07	0.61 ± 0.03	0.80 ± 0.04	0.83 ± 0.02
CONF	0.40 ± 0.05	0.44 ± 0.05	0.58 ± 0.03	0.66 ± 0.05	0.70 ± 0.02	0.82 ± 0.03
ENT	0.40 ± 0.05	0.44 ± 0.04	0.58 ± 0.03	0.66 ± 0.03	0.68 ± 0.04	0.83 ± 0.03
MCDR	0.40 ± 0.05	0.45 ± 0.03	0.65 ± 0.02	0.62 ± 0.05	0.70 ± 0.01	0.83 ± 0.03
P100	0.82 ± 0.02	0.82 ± 0.02	0.82 ± 0.02	0.82 ± 0.02	0.82 ± 0.02	0.82 ± 0.02
P99	0.81 ± 0.02	0.81 ± 0.02	0.81 ± 0.02	0.81 ± 0.02	0.81 ± 0.02	0.81 ± 0.02
Random	0.40 ± 0.05	0.63 ± 0.03	0.75 ± 0.04	0.70 ± 0.04	0.78 ± 0.04	0.82 ± 0.02

The bold font indicates the highest performance of the corresponding method

Notably, PIAA outperforms other methods, achieving 99% performance of the model trained on a fully labeled dataset while utilizing only 12% of the annotated data

Table 2 Overview of the test performance (average Dice score) is achieved by different active learning algorithms when they are combined with various deep learning architectures

Arch	p100	PIAA	CEAL	CORESET	RMCDR	MAXRPR
YN*[5]	**0.83 ± 0.02**	**0.83 ± 0.01**	0.52 ± 0.01	0.45 ± 0.02	0.44 ± 0.01	0.56 ± 0.01
YN [5]	0.82 ± 0.02	**0.81 ± 0.02**	0.48 ± 0.01	0.47 ± 0.02	0.45 ± 0.01	0.53 ± 0.01
UN[31]	0.79 ± 0.02	**0.80 ± 0.01**	0.39 ± 0.01	0.48 ± 0.02	0.63 ± 0.01	0.51 ± 0.01
DP-V3^r	0.74 ± 0.04	**0.74 ± 0.02**	0.62 ± 0.01	0.49 ± 0.01	0.57 ± 0.01	0.61 ± 0.01
DP-V3^m	0.61 ± 0.01	**0.61 ± 0.01**	0.28 ± 0.02	0.25 ± 0.01	0.59 ± 0.02	0.51 ± 0.01
DP-V3^{r,†}	0.78 ± 0.01	**0.79 ± 0.01**	0.29 ± 0.01	0.68 ± 0.01	0.68 ± 0.01	0.73 ± 0.01
DP-V3^{m,†}	0.78 ± 0.01	**0.79 ± 0.01**	0.18 ± 0.01	0.57 ± 0.01	**0.79 ± 0.02**	0.75 ± 0.02

The bold font indicates the highest performance of the corresponding method

These architectures include models with pre-trained weights and those trained with only 12% of actively selected data from the Duke dataset. The reported results (average value ± standard deviation) are obtained by averaging the outcomes from two separate runs with different random seeds. The superscript 'r' corresponds to ResNet, 'm'corresponds to MobileNet version 3 as the backbone, and '†' signifies that the networks are initialized with pre-trained weights sourced from ImageNet [32]
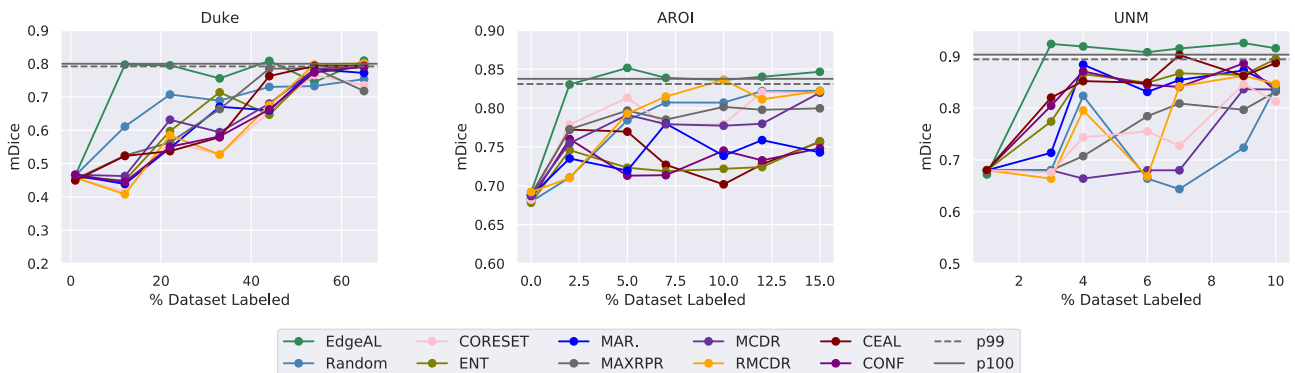

Fig. 3 The performance of the segmentation model (YN*) is indicated by the mean Dice score. PIAA and various other active learning (AL) methods are compared to baseline results across the Duke, AROI, and UNM datasets. The solid and dashed lines respectively denote 100% and 99% performance scores of the YN* when trained with the entire labeled dataset

entropy (ENT)[17], cost-effective active learning (CEAL), core-set selection (CORESET) [49, 36], and regional MC dropout entropy (RMCDR) [37], maximum representations (MAXRPR) [50], and random selection (Random).

4.1 Datasets and Networks

To evaluate the performance of our method, PIAA, we conduct experiments on three OCT segmentation datasets:
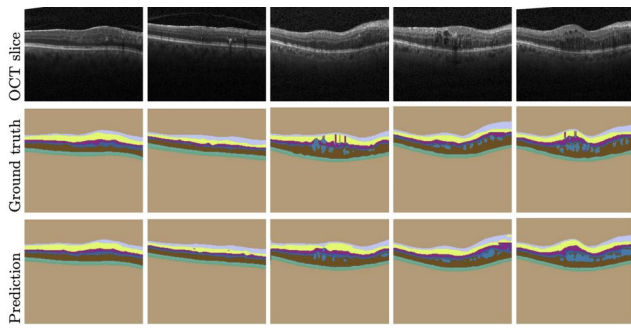
Fig. 4 Sample OCT (Duke) test images, along with manually annotated ground truth segmentation maps and our prediction results. The corresponding predictions are generated after training the model (YN*) using only 12% of the available samples

Duke [51], AROI [52], and UMN [53]. The Duke dataset consists of 100 B-scans obtained from 10 different patients, the AROI dataset contains 1136 B-scans from 24 patients, and the UMN dataset comprises 725 OCT B-scans from 29 patients. Notably, the segmentation task in these datasets involves classifying into nine, eight, and two distinct segmentation classes in Duke, AROI, and UMN, respectively, encompassing various fluid and retinal layers. In accordance with established conventions and dataset guidelines, we adhere to a 60:20:20 train-test-validation split for the experiments, ensuring that data from a single patient is not mixed across these splits.

Furthermore, for uniformity and compatibility with our experimental setup, we resize all images and their

Table 3 The fivefold test result of the YN* segmentation model trained on 12% actively selected data (**Duke**) using different active learning methods (*rows*)

Method	Above/below	ILM	NFL-IPL	INL	OPL	ONL-ISM	ISE	OS-RPE	Fluid	Overall
MAXRPR	0.82±0.08	0.12± 0.24	0.22±0.27	0.33±0.15	0.01±0.01	0.78±0.06	0.66±0.17	0.00±0.00	0.03±0.04	0.54± 0.09
MCDR	0.81±0.09	0.63±0.07	0.28±0.1	0.0±0.15	0.22±0.05	0.81±0.05	0.42±0.06	0.78±0.04	0.08±0.07	0.45±0.03
CORESET	0.77±0.10	0.67±0.04	0.23±0.08	0.0±0.00	0.21±0.23	0.82±0.05	0.45±0.09	0.75± 0.05	0.07±0.06	0.44± 0.05
CEAL	0.85±0.04	0.72±0.05	0.4±0.15	0.31±0.18	0.11±0.16	0.76±0.11	0.67±0.14	0.77±0.03	0.25±0.09	0.54± 0.04
MAR	0.72±0.10	0.67±0.08	0.2±0.06	0.00±0.00	0.14±0.12	0.76±0.10	0.48±0.13	0.77±0.05	0.24±0.05	0.44±0.05
ENT	0.73± 0.10	0.66±0.09	0.21±0.06	0.00±0.00	0.12±0.08	0.77 ±0.09	0.50±0.10	0.76±0.05	0.24±0.07	0.44±0.04
CONF	0.72±0.11	0.65± 0.09	0.21±0.08	0.00±0.00	0.15±0.11	0.76±0.09	0.48±0.12	0.77±0.05	0.23±0.09	0.44±0.05
RMCDR	0.70±0.07	0.68±0.05	0.18±0.04	0.00±0.00	0.35±0.19	0.77±0.10	0.42±0.16	0.75 ±0.06	0.07±0.05	0.44± 0.04
Random	0.98±0.01	0.60±0.70	0.78±0.04	0.61±0.05	0.05±0.05	0.84±0.05	0.72±0.06	0.79±0.03	0.28±0.12	0.63± 0.03
PIAA	**1.00±0.00**	**0.86±0.01**	**0.89±0.02**	**0.78±0.03**	**0.73±0.04**	**0.88±0.03**	**0.86±0.02**	**0.83±0.03**	**0.57±0.06**	**0.82±0.03**
p100	**0.99±0.00**	**0.86±0.01**	**0.87±0.02**	**0.71±0.11**	**0.71±0.06**	**0.86±0.06**	**0.85± 0.02**	**0.83±0.02**	**0.43±0.20**	**0.82± 0.02**

The bold font indicates the highest performance of the corresponding method

The columns represent the performance of different retinal layers and fluid layers. (Mean±sd) are calculated based on the five-fold test performances. p100 indicates the state-of-the-art test performance while the model is trained on the complete train set. From the table, we can conclude that PIAA has superior and consistent performance in fivefold cross-validation

Table 4 The class-wise (columns) performance comparison between different active learning methods (rows) on AROI dataset when YN* model is trained on 3% actively selected data

Method/class	Above ILM	ILM-IPL/INL	IPL/INL-RPE	RPE-BM	Under BM	PED	SRF	IRF	Overall
MAXRPR	1.00± 0.00	0.91± 0.01	0.85± 0.00	0.11± 0.15	0.97± 0.01	0.01± 0.01	0.31± 0.08	0.98± 0.00	0.64± 0.01
MCDR	0.98± 0.02	0.91± 0.00	0.88± 0.01	0.11± 0.03	0.98± 0.00	0.18± 0.24	0.21± 0.14	0.92± 0.09	0.65± 0.04
CORESET	1.00± 0.00	0.93± 0.01	0.90± 0.01	0.09± 0.03	0.98± 0.00	0.43± 0.01	0.37± 0.00	0.92± 0.00	0.70± 0.00
CEAL	0.99± 0.00	0.90± 0.01	0.86± 0.00	0.26± 0.18	0.98± 0.00	0.29± 0.29	0.37± 0.00	0.95± 0.04	0.70± 0.01
MAR.	0.99± 0.00	0.88± 0.02	0.87± 0.01	0.34± 0.20	0.98± 0.00	0.23± 0.19	0.37± 0.00	0.86± 0.17	0.69± 0.03
CONF	0.99± 0.00	0.88± 0.05	0.86± 0.04	0.31± 0.22	0.98± 0.00	0.18± 0.06	0.30± 0.10	0.96± 0.00	0.68± 0.02
ENT	0.99± 0.01	0.91± 0.00	0.87± 0.03	0.30± 0.27	0.98± 0.00	0.15± 0.08	0.37± 0.00	0.97± 0.00	0.69± 0.05
RMCDR	0.99± 0.01	0.93± 0.01	0.88± 0.00	0.18± 0.12	0.97± 0.00	0.10± 0.13	0.33± 0.05	0.91± 0.01	0.66± 0.01
Random	1.00± 0.00	0.92± 0.01	0.89± 0.01	0.18± 0.26	0.97± 0.02	0.18± 0.24	0.30± 0.10	0.92± 0.07	0.67± 0.05
PIAA	**1.00± 0.00**	**0.97± 0.00**	**0.95± 0.00**	**0.64± 0.01**	**0.99± 0.00**	**0.63± 0.03**	**0.52± 0.00**	**0.98± 0.00**	**0.83± 0.01**
p100	**1.00± 0.00**	**0.97± 0.00**	**0.95± 0.00**	**0.64± 0.01**	**0.98± 0.00**	**0.60± 0.06**	**0.52± 0.01**	**0.98± 0.00**	**0.83± 0.00**

The bold font indicates the highest performance of the corresponding method

Each method is trained on two random seeds, and mean Dice scores and standard deviations are reported. Columns (2–9) report the prediction results of 8 different retinal layers and fluid layers, and Column 10 (Overall) is the average score. Results show that with 3% of actively selected annotations, only our method (PIAA) achieves full training (p100) performance

Table 5 Class-wise (rows) performance comparison between different active learning methods (columns) on UMN dataset when YN* model is trained on 3% actively selected data

Class/method	MAXRPR	MCDR	CORESET	CEAL	MAR.	CONF	ENT	RMCDR	Random	PIAA	p100
Above/below	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	0.99±0.00	0.99±0.00	**1.00±0.00**	**1.00±0.00**
Fluids	0.75±0.03	0.68±0.16	0.67±0.05	0.78±0.04	0.73±0.00	0.75±0.07	0.77±0.02	0.46±0.14	0.51±0.31	**0.88±0.01**	**0.80±0.06**
Overall	0.87±0.02	0.84±0.08	0.83±0.02	0.89±0.02	0.86±0.00	0.87±0.04	0.88±0.01	0.73±0.07	0.75±0.15	**0.94±0.00**	**0.90±0.03**

The bold font indicates the highest performance of the corresponding method

Each method is trained on two random seeds, and mean Dice scores and standard deviations are reported. Results show that with 3% of actively selected annotations, only our method (PIAA) achieves full training (p100) performance

corresponding ground truth segmentations to a common resolution of 224 × 224 using a Bilinear approximation method.

To assess the robustness and generalizability of PIAA, we perform a fivefold cross-validation (CV) specifically on the Duke dataset, while taking care not to include data from specific patients in the same fold for both training and test sets. The results of this fivefold CV analysis are presented in Table 1, summarizing the performance outcomes of our approach.

We conduct experiments using the Y-net (YN) [5], U-net (UN) [31], and DeepLab-V3 (DP-V3) [39] architectures, employing both ResNet and MobileNet backbones [31]. The results of these experiments are presented in Table 2.

It's worth noting that we do not utilize any pre-trained weights in our experiments, except for the ablation study outlined in Table 2. We utilize a mixed loss combining Dice and Cross-entropy and employ the Adam optimizer with learning rates of 0.005 and a weight decay of 0.0004. The training process spans 100 epochs with a maximum batch size of 10, which remains consistent throughout all active learning iterations. Our hyperparameter settings and evaluation metric (Dice score) are in alignment with those specified in [5], which serves as the baseline for our experiments.

4.2 Comparisons

Figure 3 compares the performance of PIAA with other contemporary active learning algorithms for image annotation across three datasets. Results show that PIAA outperforms other methods on all three datasets. By using only 12% (~8 samples), 2.3% (~16 samples), and 3% (~14 samples) labeled data on Duke, AROI, and UNM datasets respectively, our method can consistently achieve 99% of maximum model performance. Other AL methods, including CEAL, RMCDR, CORESET, and MAR, require significantly more samples to achieve these performances and their performances are not consistent across the three OCT image datasets. For a fairer comparison, we report the results using the same segmentation network YN* and hyperparameters (described in Sect. 3.1) for all the active learning strategies.

Table 1 describes fivefold cross-validation results on the Duke dataset. The segmentation results are reported in *mean Dice* scores. We can observe that for all the AL methods we achieve similar performance of the segmentation model, after training on a 2% seed set. However, on 12% actively selected training data, PIAA achieves performance close to model training on full data and significantly outperforms the other AL approaches. Results in row *p100* indicate the performance on the testset when the model is trained using the complete training dataset. CEAL and MAXRPR could
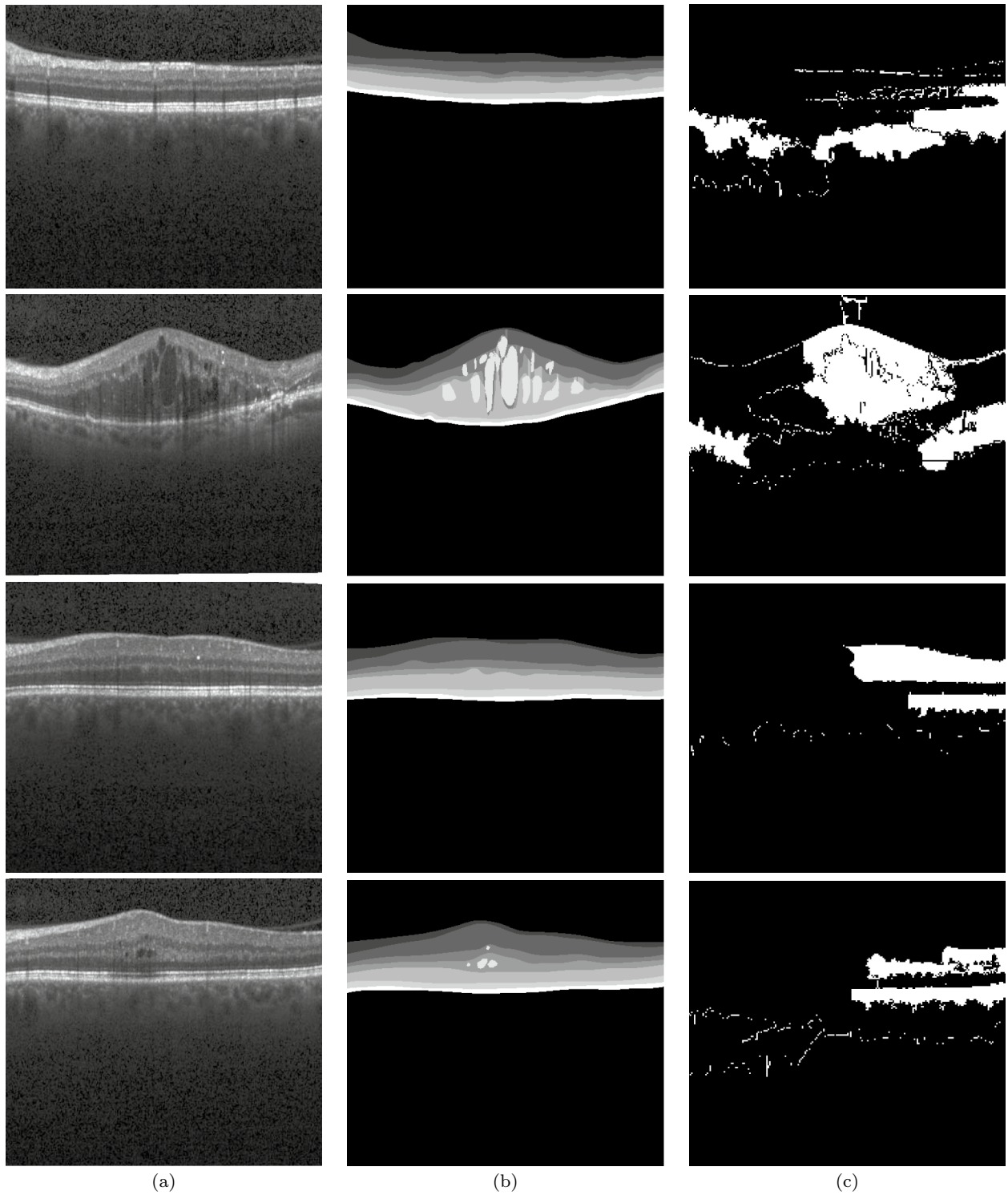
Fig. 5 Examples of **a** input images, **b** ground truth segmenting, and **c** the corresponding query (highlighted white-colored regions) from our active learning algorithm. For our method, annotation beyond the white region can be skipped during training because the model has lower uncertainty in that region. As a result, partial annotation is sufficient for training

perform similarly after training YN* on 43% of actively selected samples.

Additionally, in order to investigate the robustness of PIAA independently, we conduct experiments using four different network architectures and default weight initialization methods in PyTorch (LeCun initialization)[1] and ImageNet weight initialization. Results in Table 2 show that our proposed active learning method consistently outperforms other AL methods across different foundation models for segmentation. In contrast, other active learning methods like RMCDR and MAXRPR exhibit strong performance only when applied to pre-trained models Table 2.

Classwise segmentation performance comparison of different AL methods for different retinal and fluid layers of the Duke, AROI, and UMN datasets are also reported in Tables 3, 4 and 5 respectively. For the Duke dataset in Table 3, by training the Y-net segmentation model (YN*) with 12% actively selected data using PIAA strategy, we achieve model performance close to *p100* across all the 9 segmentation classes. The other AL methods achieve significant scores for a few classes including *ONL-ISM* and *OS-RPE* but underperform for the rest of the classes. Similar trends are also observed for different retinal and fluid layers for the AROI (Table 4) and UMN (Table 5) datasets. Moreover, Fig. 4 visually demonstrates the outputof the segmentation model trained using our AL strategy.

To highlight the significance of the partial annotation, qualitative examples are described in Fig. 5. Instead of requiring the full images to be annotated, our region-based active learning strategy, PIAA, finds the partial region that is most uncertain and needs to be annotated. Thus reducing the annotation effort significantly.

5 Conclusion

PIAA is a novel active learning technique for OCT image segmentation, which accomplished results similar to full training with a relatively small amount of data by utilizing edge information to identify regions of uncertainty. By harnessing edge information, our method not only minimized the labeling effort but also exhibited significant promise in the medical domain where labeled data is scarce. The performance of PIAA in OCT segmentation suggests that a significant amount of data is not always required to learn data distribution in medical imaging.

Furthermore, the reliance on edges as a fundamental image characteristic positions PIAA for potential adaptation and application in diverse domains with minimal model modifications. This adaptability opens up prospects for future research and development, encouraging exploration in other classical image processing and analysis areas. Our findings suggest that PIAA can significantly contribute to future active learning techniques for more resource-efficient and effective methods across multiple fields.

Our future work in the development of our active learning framework will focus on enhancing scalability, security, and efficiency while maintaining data sovereignty. Maintaining the privacy and security of patient data is paramount. To address these concerns, we have planned to implement the Minimum Data Transfer Service to anonymize structured and unstructured data, including medical images. This should ensure that no identifiable patient information transfers to any Cloud Computing Services by minimizing training data, and aligning with data sovereignty requirements.

Storing the entire medical data on the cloud can lead to privacy breaches, regulatory non-compliance, data sovereignty issues, increased security risks, and higher operational costs. The proposed PIAA active learning strategy will be utilized for selectively identifying and labeling relevant data, minimizing unnecessary data exposure to the cloud. Currently, the active learning framework is deployed on GPUs with limited power. We plan to transition the model training step to the cloud platform. The use of Sovereign Cloud may provide additional data protection measures compared to standard cloud platforms and could enhance trust among healthcare industry partners for the hosting of healthcare data in a secure manner.

---

[1] https://pytorch.org

References

1. Yuan W, Lu D, Wei D, Ning M, Zheng Y (2022) Multiscale unsupervised retinal edema area segmentation in OCT images. In: Medical image computing and computer assisted intervention-MICCAI, 25th international conference, September 18–22, 2022, proceedings, part II. Springer, Singapore, pp 667–676

2. Sonntag D (2019) Medical and health systems. In: Oviatt SL, Schuller BW, Cohen PR, Sonntag D, Potamianos G, Krüger A (eds) The handbook of multimodal-multisensor interfaces: language processing, software, commercialization, and emerging directions-vol 3. Association for Computing Machinery. https://doi.org/10.1145/3233795.3233808

3. Nath V, Yang D, Roth HR, 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, 25th international conference, September 18–22, 2022, proceedings, part VIII. Springer, Singapore, pp 297–308

4. Nguyen DMH, Ezema A, Nunnari F, Sonntag D (2020) A visually explainable learning system for skin lesion detection using multiscale input with attention U-Net. In: KI 2020: advances in artificial intelligence: 43rd German conference on AI, Bamberg, Germany, September 21–25, 2020, proceedings 43. Springer, pp 313–319

5. Farshad A, Yeganeh Y, Gehlbach P, Navab N (2022) Y-Net: a spatiospectral dual-encoder network for medical image segmentation. In: Medical image computing and computer assisted intervention-MICCAI, 25th international conference, September 18–22, 2022, proceedings, part II. Springer, Singapore, pp 582–592

6. Tusfiqur HM, Nguyen DM, Truong MT, Nguyen TA, Nguyen BT, Barz M, Profitlich H-J, Than NT, Le N, Xie P et al (2022) Drg-net: interactive joint learning of multi-lesion segmentation and classification for diabetic retinopathy grading. arXiv preprint arXiv:2212.14615

7. Lu F, Tang C, Liu T, Zhang Z, Li L (2023) Multi-attention segmentation networks combined with the sobel operator for medical images. Sensors 23(5):2546

8. Kadir MA, Alam HMT, Sonntag D (2023) EdgeAL: an edge estimation based active learning approach for OCT segmentation. In: International conference on medical image computing and computer-assisted intervention. Springer, pp 79–89

9. Lee B, Paeng K (2018) A robust and effective approach towards accurate metastasis detection and Pn-stage classification in breast cancer. In: Medical image computing and computer assisted intervention-MICCAI, 21st international conference, September 16–20, 2018, proceedings, part II 11. Springer, Granada, pp 841–850

10. Samrath S, Sayna E, Trevor D (2019) Variational adversarial active learning. In: 2019 IEEE/CVF international conference on computer vision (ICCV). IEEE

11. Dai C, Wang S, Mo Y, Zhou K, Angelini E, Guo Y, Bai W (2020) Suggestive annotation of brain tumour images with gradient-guided sampling. In: Medical image computing and computer assisted intervention-MICCAI, 23rd international conference, October 4–8, 2020, proceedings, part IV 23. Springer, Lima, Peru, pp 156–165

12. Bai F, Xing X, Shen Y, Ma H, Meng MQ-H(2022) Discrepancy-based active learning for weakly supervised bleeding segmentation in wireless capsule endoscopy images. In: Medical image computing and computer assisted intervention-MICCAI, 25th international conference, September 18–22, 2022, proceedings, part VIII. Springer, Singapore, pp 24–34

13. Lewis DD, Catlett J (1994) Heterogeneous Uncertainty Sampling for Supervised Learning. In: Machine learning proceedings. Elsevier, pp 148–156

14. Nath V, Yang D, Landman BA, Xu D, Roth HR (2020) Diminishing uncertainty within the training pool: active learning for medical image segmentation. IEEE Trans Med Imaging 40(10):2534–2547

15. Joshi AJ, Porikli F, Papanikolopoulos N (2009) Multi-class active learning for image classification. In: IEEE conference on computer vision and pattern recognition. IEEE, pp 2372–2379

16. Hwa R (2004) Sample Selection for Statistical Parsing. Comput Linguist 30(3):253–276

17. Luo W, Schwing A, Urtasun R (2013) Latent structured active learning. In: Advances in neural information processing systems, vol 26, New York, pp 728–736

18. Balaram S, Nguyen CM, Kassim A, Krishnaswamy P (2022) Consistency-based semi-supervised evidential active learning for diagnostic radiograph classification. In: Medical image computing and computer assisted intervention-MICCAI, 25th international conference, September 18–22, 2022, proceedings, part I. Springer, Singapore, pp 675–685

19. Yoo D, Kweon IS (2019) Learning loss for active learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), California, pp 93–102

20. Lewis DD (1995) A sequential algorithm for training text classifiers: corrigendum and additional data. In: Acm Sigir Forum, vol 29. ACM, New York, pp 13–19

21. Nguyen HT, Smeulders A (2004) Active learning using pre-clustering. In: Proceedings of the twenty-first international conference on machine learning, New York, pp 79

22. Yang Y, Ma Z, Nie F, Chang X, Hauptmann AG (2015) Multi-class active learning by uncertainty sampling with diversity maximization. Int J Comput Vis 113:113–127

23. Guo Y (2010) Active instance sampling via matrix partition. In: Advances in neural information processing systems, vol 23, Vancouver, pp 802–810

24. Elhamifar E, Sapiro G, Yang A, Sasrty SS (2013) A convex optimization framework for active learning. In: Proceedings of the IEEE international conference on computer vision, Sydney, pp 209–216

25. Bilgic M, Getoor L (2009) Link-based active learning. In: NIPS workshop on analyzing networks and learning with graphs, vol 4. pp 9

26. Mac Aodha O, Campbell ND, Kautz J, Brostow GJ (2014) Hierarchical subquery evaluation for active learning on a graph. In: Proceedings of the IEEE conference on computer vision and pattern recognition, Ohio, pp 564–571

27. Settles B, Craven M, Ray S (2007) Multiple-instance active learning. In: Advances in neural information processing systems, vol 20. Vancouver, pp 1289–1296

28. Roy N, McCallum A (2001) Toward optimal active learning through monte carlo estimation of error reduction, vol 2. ICML, Williamstown, pp 441–448

29. Freytag A, Rodner E, Denzler J (2014) Selecting influential examples: active learning with expected model output changes. In: Computer vision-ECCV, 13th European conference, September 6–12, 2014, proceedings, part IV 13. Springer, Zurich, Switzerland, pp 562–577

30. Huang H, Lin L, Tong R, Hu H, Zhang Q, Iwamoto Y, Han X, Chen Y-W, Wu J (2020) Unet 3+: a full-scale connected Unet for medical image segmentation. In: ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 1055–1059

31. Khan A, Sohail A, Zahoora U, Qureshi AS (2020) A survey of the recent architectures of deep convolutional neural networks. Artif Intell Rev 53:5455–5516

32. Deng J, Dong W, Socher R, Li L-J, Li K, Li F-F (2009) Imagenet: a large-scale hierarchical image database. 2009 IEEE conference on computer vision and pattern recognition, Florida, pp 248–255

33. Lin L, Wang K, Meng D, Zuo W, Zhang L (2017) Active self-paced learning for cost-effective and progressive face identification. IEEE Trans Pattern Anal Mach Intell 40(1):7–19

34. Gal Y, Islam R, Ghahramani Z (2017) Deep Bayesian active learning with image Dta. In: International conference on machine learning. PMLR, pp 1183–1192

35. Beluch WH, Genewein T, Nürnberger A, Köhler JM (2018) The power of ensembles for active learning in image classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, Utah, pp 9368–9377

36. Gorriz M, Carlier A, Faure E, Giro-i Nieto X (2017) Cost-effective active learning for melanoma segmentation. arXiv preprint arXiv:1711.09168

37. Mackowiak R, Lenz P, Ghori O, Diego F, Lange O, Rother C (2018) Cereals-cost-effective region-based active learning for semantic segmentation. arXiv preprint arXiv:1810.09726

38. Wang K, Zhang D, Li Y, Zhang R, Lin L (2016) Cost-effective active learning for deep image classification. IEEE Trans Circuits Syst Video Technol 27(12):2591–2600

39. Siddiqui Y, Valentin J, Nießner M (2020) ViewAL: Active learning with viewpoint entropy for semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, Washington, pp 9433–9443

40. Muslea I, Minton S, Knoblock CA (2006) Active learning with multiple views. J Artif Intell Res 27:203–233

41. Al-Amri SS, Kalyankar N, Khamitkar S (2010) Image segmentation by using edge detection. Int J Comput Sci Eng 2(3):804–807

42. Li X, Niu S, Gao X, Liu T, Dong J (2021) Unsupervised domain adaptation with self-selected active learning for cross-domain OCT image segmentation. In: Neural information processing: 28th international conference, ICONIP (2021) Sanur, December 8–12, 2021, proceedings, Part II 28. Springer, Bali, Indonesia, pp 585–596

43. Chi L, Jiang B, Mu Y (2020) Fast Fourier convolution. Adv Neural Inf Process Syst 33:4479–4488

44. Ronneberger O, Fischer P, Brox T (2015) U-net: convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention-MICCAI, (2015) 18th international conference, October 5–9, 2015, proceedings, part III 18. Springer, Munich, Germany, pp 234–241

45. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, Desmaison A, Kopf A, Yang E, DeVito Z, Raison M, Tejani A, Chilamkurthy S, Steiner B, Fang L, Bai J, Chintala S (2019) PyTorch: an imperative style, high-performance deep learning library. In: Advances in neural information processing systems 32. Curran Associates, Inc., pp 8024–8035. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

46. Zhao Z, Wallace E, Feng S, Klein D, Singh S (2021) Calibrate before use: improving few-shot performance of language models. In: Meila M, Zhang T (eds) Proceedings of the 38th international conference on machine learning, ser. Proceedings of machine learning research, vol 139. PMLR, pp 12697–12706

47. Yun S, Park J, Lee K, Shin J (2020) Regularizing class-wise predictions via self-knowledge distillation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), Washington, pp 13876–13885

48. Van den Bergh M, Boix X, Roig G, de Capitani B, Van Gool L (2012) SEEDS: superpixels extracted via energy-driven sampling. ECCV 7(7578):13–26

49. Sener O, Savarese S (2017) Active learning for convolutional neural networks: a core-set approach. arXiv preprint arXiv:1708.00489

50. Yang L, Zhang Y, Chen J, Zhang S, Chen DZ (2017) Suggestive annotation: a deep active learning framework for biomedical image segmentation. In: Medical image computing and computer assisted intervention-MICCAI, 20th international conference, September 11–13, 2017, proceedings, part III 20. Springer, Quebec City, QC, Canada, pp 399–407

51. Chiu SJ, Allingham MJ, Mettu PS, Cousins SW, Izatt JA, Farsiu S (2015) Kernel regression based segmentation of optical coherence tomography images with diabetic macular edema. Biomed Opt Express 6(4):1172–1194

52. Melinščak M, Radmilovč M, Vatavuk Z, Lončarić S (2021) AROI: annotated retinal OCT images database. In: 44th international convention on information. Communication and electronic technology (MIPRO), Croatia, pp 371–376

53. Rashno A, Nazari B, Koozekanani DD, Drayna PM, Sadri S, Rabbani H, Parhi KK (2017) Fully-automated segmentation of fluid regions in exudative age-related macular degeneration subjects: kernel graph cut in neutrosophic domain. PLoS One 12(10):e0186949