Lappeenranta-Lahti University of Technology LUT
School of Engineering Science
Computational Engineering and Technical Physics
Computer Vision and Pattern Recognition

Tuo Yang

HUMAN-IN-THE-LOOP FOR EFFICIENT TRAINING OF RETINAL IMAGE ANALYSIS METHODS

Master's Thesis

Examiners: Professor Lasse Lensu
Professor Vyacheslav Potekhin

Supervisors: M.Sc.(Tech.) Sinan Kaplan
Professor Lasse Lensu

ABSTRACT

Lappeenranta-Lahti University of Technology LUT
School of Engineering Science
Computational Engineering and Technical Physics
Computer Vision and Pattern Recognition

Tuo Yang

Human-in-the-loop for efficient training of retinal image analysis methods

Master's Thesis

2021

80 pages, 24 figures, 17 formulas.

Examiners: Professor Lasse Lensu
Professor Vyacheslav Potekhin

Keywords: computer vision, machine vision, retinal image analysis, active learning

Eye-related diseases like diabetic retinopathy can be visually diagnosed by medical experts. However, this is time-consuming work requires much effort. For solving this problem, automated computer-aided solutions have been proposed based on retinal image analysis methods. However, this kind of system generally requires lots of data annotated by experts for training the methods for getting relevant analysis results. For the efficient utilization of expert's time, active learning can be used to achieve good enough image analysis results. The experimental part of the work studies active learning strategies for the segmentation of retinal blood vessels, then compare the results to a benchmark. Based on the experiments, most active learning strategies show better segmentation accuracy than random sampling as the benchmark. In addition, when the number of images available for the training is increased, active learning performs well against the fully supervised learning model.

PREFACE

I am greatly indebted to my supervisor, Professor Lasse Lensu, for his valuable instructions and suggestions on my thesis. Also, with his constant support and encouragement, I am successfully able to complete this thesis work. Meanwhile, I would also like to extend my appreciation to Sinan Kaplan for his advice during my research.

Last but not least, I would express my heartfelt gratitude to my friends and classmates for their valuable advice and cooperation throughout the preparation of the thesis work.

Lappeenranta, June 1, 2021

Tuo Yang

CONTENTS

LIST OF ABBREVIATIONS

A/V	Artery/Vein
AHE	Adaptive Histogram Equalization
AL	Active learning
AMD	Adaptive Maximize Disagree
AOO	American Academy of Ophthalmology
AVR	The Arteriolar-to-Venular diameter Ratio
BMDAL	Batch Model Deep Active Learning
CDF	Cumulative Distribution Function
CGAN	Conditional Generative Adversarial Network
CLAHE	Contrast Limited Adaptive Histogram Equalization
CNN	Convolutional Neural Network
CRF	Conditional Random Field
DAL	Deep Active Learning
DBAL	Deep Bayesian Active Learning
DNN	Deep Neural Network
EER	Excepted Error Reduction
EMC	Excepted Model Change
EQB	Entropy Query-by-Bagging
ETDRS	Early Diabetic Retinal Diagnostic Society
FOV	Field of View
GAN	Generative Adversarial Network
GT	Ground Truth
HE	Histogram Equalization
KNN	K-Nearest Neighbor
LC	Least Confidence
MCLU	Multi-Class Level Uncertainty
MRI	Magnetic Resonance Imaging
MS	Margin Sampling
NAS	Neural Architecture Search
NDPR	Nonproliferative Diabetic Retinopathy
PA	Pixel Accuracy
PDR	Proliferative Diabetic Retinopathy
QBC	Query By Committee
ROI	Region of Interest
SVM	Support Vector Machine
TNR	True Negative Rate

TPR	True Positive Rate
US	Uncertainty Sampling

1 INTRODUCTION

1.1 Background

With the multiple health threats increasing worldwide, eye diseases, as one of them, have gradually caught people's attention. For instance, diabetic retinopathy, diabetic nephropathy, and diabetic neuropathy are three major types of diabetic abnormalities [1]. Among these diseases, medical experts can judge eye diseases like diabetic retinopathy from eye fundus images. The retinal image analysis is generally used for making the diagnosis for diabetic retinopathy.

The development of retinal image analysis methods has several stages. The first one is early-stage retinal image analysis, which uses image processing algorithms to do detection and segmentation for corresponding retinopathy lesions [2]. However, with the tremendous success of the deep learning algorithms in the image classification field, more researchers gradually adapt them to realize classification [3] and lesion detection [4] to fundus images of diabetic retinopathy as well. Its performance exceeds that of the traditional diabetic retinopathy detection algorithm. In [5], there are three typical retinal image analyses: retinopathy degree classification, retinal image blood vessel segmentation, and red lesion detection. For getting the best image analysis results, the realization of these methods can be done by training a Convolutional Neural Network (CNN) model using supervised learning.

The performance of CNN is better than traditional image processing algorithms, especially on retinal lesion classification and retinal lesion detection [3, 4]. It does not require professional knowledge to design lesion features. However, based on the CNN model, retinal image analysis methods still need many annotated trained samples made by experts, which will generate a high cost of making annotations and restrict the CNN applications.

For the problems mentioned above, Human in the loop is a feasible solution. It mainly relies on human intelligence to help the machine to become more intelligent. Annotation and Active learning (AL) are the cornerstones of Human-in-the-Loop Machine Learning [6]. AL is a subfield of machine learning, or more generally, artificial intelligence. There are two important modules of the AL algorithm: learner and selection strategy. AL uses the selection strategy, actively select some samples from the unlabeled sample set, then provide them for experts in related fields to label, lastly add the labeled samples to

the training data set for the learning module to train. The program stops when the learning module meets the termination conditions. Otherwise, the above steps are repeated continuously to obtain more labeled samples for training [7].

Some great work of active learning has shown up in the recent few years. In [8], authors propose an algorithm based on Active Learning + Transfer Learning, Data Augmentation, Majority Selection, Continuously Fine-Tuning, and other methods. Experimental results have verified that active selection strategies (entropy + diversity) can reduce at least half of the data labeling cost in three medical image sets. In [9], a combination of Generative Adversarial Network (GAN) and Active Learning has been proposed for the first time. It obtains a generator model by training the GAN, actively generates the most valuable samples for experts to mark. In [10], Konyushkova et al. proposed learning active learning from data, which is fundamentally different from the traditional active selection strategy. It overcomes the shortcomings of manual design selection strategy cross-domain generalization ability, learns by transforming active selection strategy into a regression problem. The learned strategies have achieved significant effects on real data sets (Striatum, Magnetic Resonance Imaging (MRI), Credit Card, Splice, and Higgs) in many various fields.

1.2 Goals and delimitations

This thesis concentrates on designing an effective active learning algorithm to train retinal image analysis methods, thereby selecting the most valuable samples from unlabeled data for an expert to make annotations, thus improving the performance of retinal image analysis methods at the minimal cost of making annotations.

More specifically, the research goals are as follows:

- Study retinal image content from the viewpoint of eye diseases and relevant literature on active learning for retinal image analysis.
- Study and implement one or more active learning approaches, then use them to select the next image to be annotated, thus improving retinal image analysis performance.
- Quantitatively evaluate the learning against a benchmark.

1.3 Structure of the thesis

The whole thesis architecture is arranged as follows: Chapter 2 introduces background knowledge related to the retinal image and its corresponding analysis methods. This chapter also introduces the related research on active learning, including its scenarios, selecting/querying strategies for unlabeled samples, and active learning applications on efficiently training for retinal image analysis methods. In Chapter 3, one type of active learning algorithm, based on the Deep Neural Network (DNN) model, is proposed to train the retinal image analysis methods efficiently. Chapter 4 describes experimental results after implementing the proposed algorithm, then evaluating the algorithm performance as well. Chapter 5 discusses the shortage of the realized algorithm, thus discussing improvements to the proposed algorithm. Finally, Chapter 6 gives one general conclusion to the whole finished work.

2 RETINAL IMAGES ANALYSIS AND ACTIVE LEARNING

This chapter mainly introduces a basic understanding of the retina, common retinal abnormalities, image analysis methods, common databases of retinal images, and active learning methods applied in the medical field to analyze the retinal image for medical diagnosis.

2.1 Physiological structure of retina

2.1.1 Structure of eyeball and retina

The human eye is approximately spherical, and its main components include the iris, pupil, ciliary muscle, lens, retina, and optic nerve [11]. The anatomical results are shown in Figure 1 [12]. The structure of the human eye is an optical imaging system. The light is transmitted to the inside of the eyeball through the pupil, forming the target's influence on the retina, then transmitting the visual information through the optic nerve to the next signal processing station.

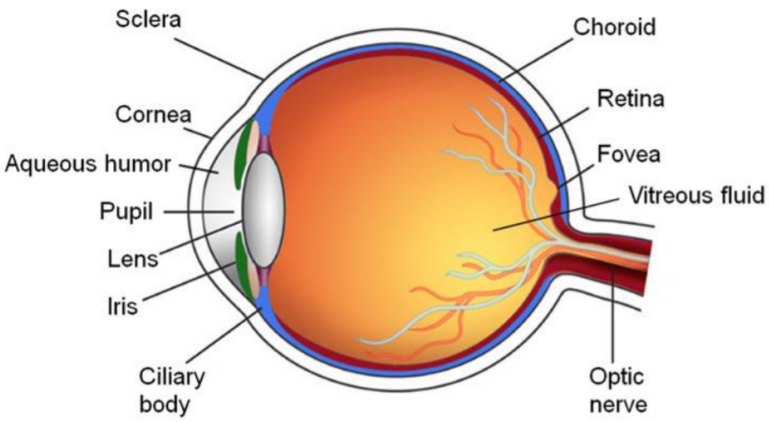


Figure 1. Schematic diagram of eyeball anatomy results. [12]

The retina is the most complex part of the fundus structure, and it is composed of important structures such as the retina, fovea, and optic disk. The structure of the retinal is presented in Figure 2 [11].

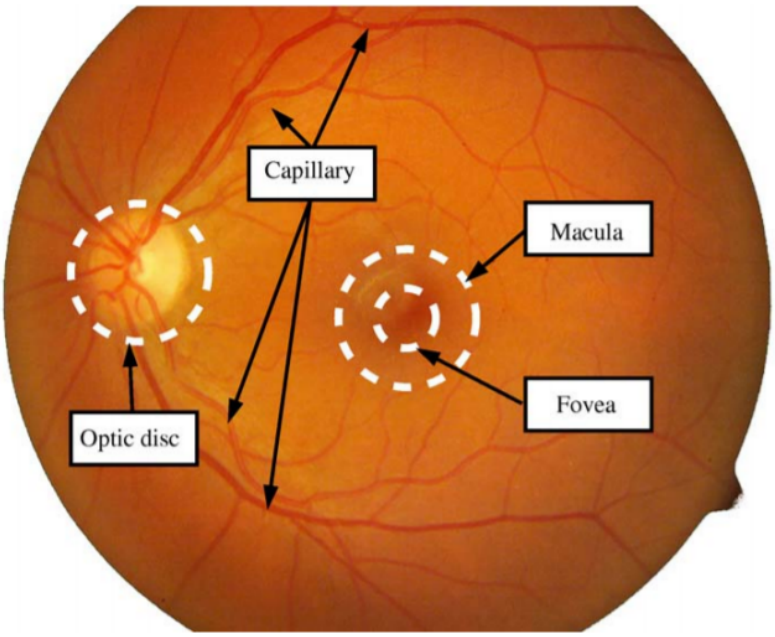


Figure 2. Schematic diagram of eyeball anatomy results. [11]

Macula [13]: an elliptical bright spot with a diameter of about $1.5\,\mathrm{mm}$, about $3.5\,\mathrm{mm}$ from the temporal edge of the optic disk, $0.3\,\mathrm{mm}$ below the horizontal meridian. The pigment content in the macula's pigment epithelial cells is higher, so the color observed under the mirror is darker. Many cone cells accumulate in the macula, accounting for about 10% of the retina's total number of cone cells. The visual acuity is the highest, and the color sensitivity decreases when the distance from the macula is farther.

Fovea [13]: a small depression with a diameter of $0.1\,\mathrm{mm}$ in the center of the macula. There are blood vessels and only very sharp cone cells. If the light reflection point is observed at this place with the ophthalmoscope, it is the macula's foveal light reflection.

Optic disk [13]: a light red area with a diameter of $1.5\,\mathrm{mm}$, located about $3\,\mathrm{mm}$ on the nasal side of the macula, has no photoreceptor cells and presents an inherent dark area in the visual field, which is also called a physiological blind spot.

Besides, the retinal tissue is also distributed with capillaries that can be directly observed in the human body's organs. These vascular diseases can reflect the human body's thrombosis, diabetes, hypertension, and other diseases [13].

2.1.2 Common types of retinal abnormalities

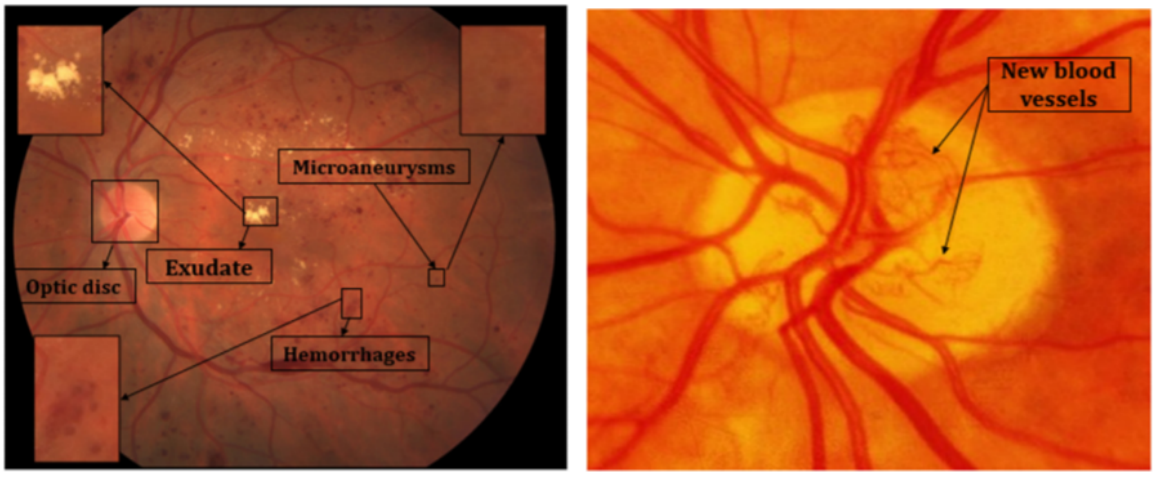
Common types of retinopathy are shown in Figure 3 [14], Figure 3 [14] shows common pathologies.These include Microaneurysms, Exudates, Hemorrhages, and New blood vessel routes. The main features are as follows [14]:

1. **Microaneurysms**: generally, small red round spots appear on the retinal image of the fundus, and their size is smaller than the diameter of blood vessels near the optic disk, which can reflect the early changes of blood vessels to a certain extent, as shown in Figure 3 (a).

2. **Exudate**: due to the continuous expansion of retinal blood vessels, the rupture of nutrients such as lipids and proteins appears on the retina. It appears bright white on the color image of the retina at the fundus. The shape is generally plaques with different sizes, as shown in Figure 3 (a).

3. **Hemorrhages**: the blood vessels in the retina rupture due to the continuous expansion of blood vessels in the retina, and the color image of the retina at the fundus appears dark red, as shown in Figure 3 (a).

4. **Blood vessel routes**: blood vessel occlusion leads to ischemia, causing the retina to produce small disordered new blood vessels for scarce blood, as shown in Figure 3 (b).



(a) (b)

Figure 3. Common types of retinopathy: (a) Microaneurysms, Exudate and Hemorrhages; (b) new blood vessel routes. [14]

The grade of the retina's disease can be divided according to different factors such as the type and location of the disease. Taking diabetic retinopathy as an example, there are different evaluation standards according to the actual application scenarios: two classifications (normal and abnormal), three classifications (normal, Nonproliferative Diabetic Retinopathy (NDPR), Proliferative Diabetic Retinopathy (PDR)), four classifications (normal, mild NDPR, severe NDPR, PDR), five categories (normal, mild NDPR, moderate NDPR, severe NDPR, PDR). The Early Diabetic Retinal Diagnostic Society (ETDRS) in 1991 proposed a worldwide uniform classification standard for diabetic retinopathy [15], which was initiated by the American Academy of Ophthalmology (AOO) [16] in 2001, and was more practical in February 2003 The global diabetic retina classification standard. The specific manifestation of the standard is shown in Figure 4 [17].

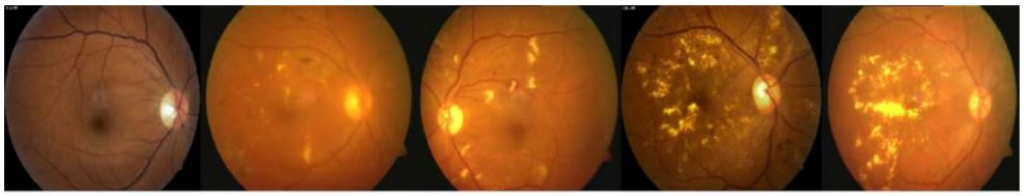


Figure 4. Different degrees of diabetic retinopathy
(left to right: normal, mild, moderate, severe, PDR). [17]

2.2 Common datasets of fundus images

The data set of fundus image lesion analysis, as one of the hot research fields, is opened to compare the pros and cons of various algorithms. Nine public fundus image datasets are introduced below, which are popular and have a significant role in promoting this field development.

1. **STARE data set** [18] has 20 pictures for blood vessel segmentation, ten of which contain Lesions. Each picture resolution is 605*700, and a Topcon TRV-50 color took it with a 35-degree field of view. Usually, each picture is marked by two corresponding markers. The first marker will mark the pictures with 32,000 pixels as blood vessels on average, while the second marker will need to mark 46,100 pixels. The first marker is better than the second marker on the edges of blood vessels and fine blood vessels, which is usually used as a fundamental ground truth.

2. **DRIVE data set** [19] is used to segment blood vessels. The data comes from a screening project for diabetic retinopathy in the Netherlands in 2004. The entire database con-

tains 40 pictures in jpg format, having a resolution of 768*584, seven of which contain Lesions. Each picture contains fine markings of the location of blood vessels without any abnormal structure markings. The training set and the test set each have 20 pictures. There are two labeling results for the test set. The first labeling result contains 12.7% of the vascular structure, and the second labeling result contains 12.3% of the vascular structure.

3. **Standard Diabetic Retinopathy Database** [20, 21] includes two sub-datasets, DIARETDB0 and DIARETDB1, published in 2006 and 2007 respectively. DIARETDB0 has 130 pictures, 110 of which are lesion pictures, and 20 are normal pictures. The lesion's specific information is only a text description, and there is no specific location mark. DIARETDB1 contains 89 pictures, of which only five are normal. The remaining lesion pictures include pixel-level annotations for microaneurysms, hemorrhages, hard or soft exudates. The resolution of the picture is usually 1500*1152.

4. **Messidor data set** [22] contains a total of 1200 pictures, taken with a Topcon TRC NW6 color camera at a 45-degree angle. There are three scale pictures: 1440*960, 2240*1488, 2304*1536. Usually, there are two markers for each picture in the database: the lesion grade and the severity of macular edema. The lesion grade is divided according to new blood vessel routes and the number of microaneurysms and hemorrhages. It is usually divided into four lesion grades, including the average category, and macular edema is divided into three grades, according to the distance between the exudation location and the macula.

5. **REVIEW data set** [23] contains a total of 16 mydriatic pictures, and each picture has a blood vessel label. This data set is mainly used to estimate the width of the blood vessel and the analysis of the structure of the blood vessel junction.

6. **ROC data set** [24] comes from a part of the online microaneurysm competition held by the University of Northern Iowa in 2009. A total of 50 pixel-level annotated pictures are released in this game, with a total of three scale sizes, namely 768*576, 1058*1061, and 1389*1383.

7. **HEI-MED data set** [25] contains a total of 169 pictures, of which 7 pictures contain microaneurysms, 54 pictures contain exudation, 73 pictures contain cotton patch, soft exudation, and other lesions. Each picture contains pixel-level markers and its resolution is 2196*1958.

8. **e_optha data set** [26] contains two data subsets, e_optha EX and e_optha MA. e_optha EX contains 35 normal pictures and 47 pictures with oozing. There are four sorts of pictures with different sizes, ranging from 1440*960 to 2544*1696. In the picture, there

are also normal structures such as reflections that significantly interfere with exudation detection.e_optha MA contains 233 pictures without microaneurysms and 148 pictures of microaneurysms with pixel-level markers. Both exudates and microaneurysms have pixel-level markings.

9. **"Diabetic Retinopathy Detection" Competition Data Set in Kaggle** [27] is commonly used for 5-class retinopathy classification, which contains 35,126 training pictures and 53,576 test pictures. Each picture will have a grade label (From 0 to 4). The picture data in this dataset comes from pictures taken by different fundus cameras.

10. **CHASEDB1** [28] includes 28 retinal images taken from the eyes of 14 school children. Usually, the first 20 images are used for training, and the remaining eight images are used for testing. The size of each image is 999×960, and the binary Field of View (FOV) mask and segmentation ground truth are obtained by manual methods.

11. **HRF** [29] contains images of 15 healthy patients, 15 images of diabetic retinopathy patients, and 15 images of glaucoma patients. Each image has a binary standard label image of blood vessel segmentation. FOV is also provided for specific data sets. The standard label data is generated by a team of experts in retinal image analysis and clinicians in cooperating ophthalmology clinics. The size of each image is 3504×2336.

2.3 Retinal image analysis methods

This section introduces retinal image analysis methods from two aspects: retinal blood vessel segmentation and classification.

2.3.1 Retinopathy blood vessel segmentation

Retinal blood vessel segmentation is to train the corresponding model (e.g., CNN), which uses the ground truth label information of the blood vessel pixel level. Thus the model can achieve the effect of automatically segmenting the blood vessel of the input fundus image. For example, a fundus image and its corresponding segmented binary graph is shown in Figure 5 [30].

In general, the existing blood vessel segmentation algorithms can be divided into supervised learning and unsupervised learning. Supervised learning algorithms need to have the vascular binary map labels marked by experts as samples, then extract each pixel's

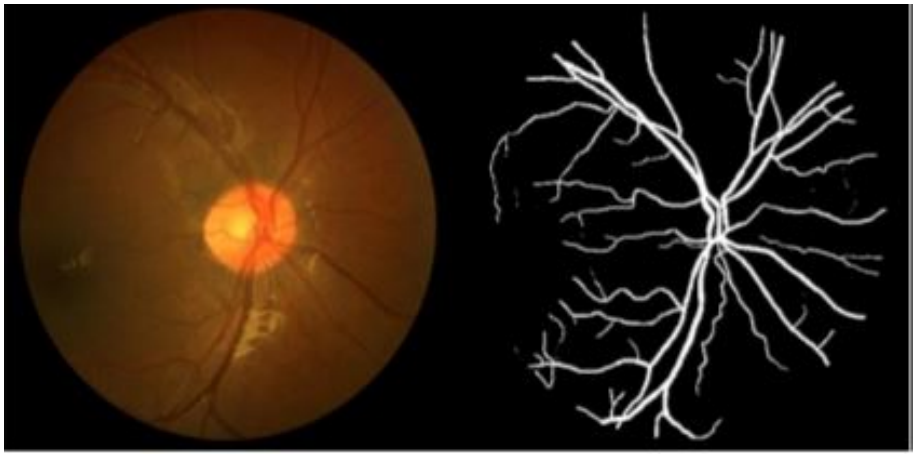


Figure 5. Retinal blood vessel segmentation results
(left: fundus image right: corresponding segmentation results). [30]

features and labels to train the classifier. Common classifiers include Support Vector Machine (SVM) [31], Conditional Random Field (CRF) [32], and CNN [30, 33–35]. For unsupervised learning, blood vessel label samples are unnecessary. However, for evaluating the results of unsupervised segmentation, the Ground Truth (GT) is needed. Most unsupervised learning algorithms are mainly based on matched filtering, mathematical morphology, blood vessel tracking, and regional growth.

The essence of using supervised learning to perform blood vessel segmentation is to classify pixels. The main purpose is to segment blood vessels using the labels, marked by experts as training samples for model training. The most important step in supervised learning is to extract the feature vector, which distinguishes blood vessels from non-vessels from the training data. Generally, the segmentation accuracy of supervised learning algorithms is higher than that of unsupervised learning.

To improve the segmentation accuracy of diseased blood vessels, Strisciuglio et al. [36] used a set of selective BCOSFIRE filters to extract different features of retinopathy images through these filters, thus achieving the effect of segmenting blood vessels in retinopathy images. Ganjee et al. [37] used a method, which is a combination of regional features and multi-scale matched filtering, distinguishing lesion structures from blood vessels.

A blood vessel segmentation algorithm based on a decision tree classifier was proposed by Fraz et al. [38]. The main principle is to train the decision number classifier by extracting the line feature, direction feature, and the gradient vector field's morphological feature.

Conditional Random Field was first introduced into blood vessel segmentation by Orlando et al. [32]. Through the support vector machine with structured output, the model

parameters are continuously learned, thus achieving the purpose of training a fully connected conditional random field (Full connected CRF) model, and finally solve the error caused by the lack of prior knowledge of the slender structure of blood vessels.

With the successful application of deep learning algorithms in computer vision, CNN has become the most successful medical image segmentation algorithm. In this context, a blood vessel classification algorithm, which combines random forests classifier and CNN, was proposed by Wang et al. [33]. After that, the algorithm of using the deep autoencoder model, constructing the relationship between retinopathy images and blood vessels, was proposed by Li et al. [30], which has a better segmentation effect for retinopathy images and microvessels. Laskowski et al. [34] used a deep neural network based on image blocks to segment blood vessels, which exceeded all previous algorithms in the blood vessel segmentation of retinopathy images. Fu [35] regards the problem of blood vessel segmentation as edge detection, first uses multi-level and multi-scale CNN to extract the retinopathy image features, and then combines CNN and random conditions to segment the retinal image blood vessels.

Unsupervised learning is rule-based learning without any prior knowledge and labeled samples. The main blood vessel extraction algorithms mainly use model-based methods and filter response methods. According to the different image processing methods, algorithms can be subdivided into matched filtering methods, blood vessel tracking methods, and model-based methods.

The matched filter uses a two-dimensional convolution kernel to convolve the retinopathy image. When there are blood vessels, the matched filter has a high response. Kovacs et al. [39] used a self-correcting algorithm based on contour reconstruction and template matching for blood vessel segmentation. Azzopardi et al. [40] used a series of translation filter response (COSFIRE) combinations to extract blood vessels. Although the matched filtering algorithm has a better segmentation effect on healthy retina images, it has a higher probability of false positives when used on fundus images with retinopathy.

The blood vessel tracking method [41] uses separating the blood vessel between two points. That is, it relies on the seed point in the local area to detect the blood vessel. The gray intensity and curvature degree determine the center of the blood vessel's longitudinal section. This method can effectively calculate the blood vessel's accurate width, but if seed point placement is not accurate, it will cause the blood vessel without the seed point to be detected.

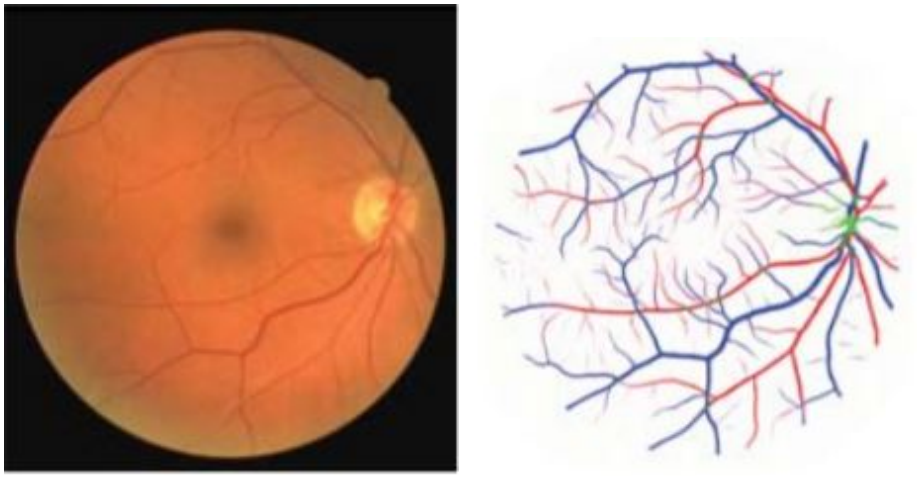
Model-based methods usually use explicit blood vessel models, such as blood vessel con-

tour models and deformation models, to extract blood vessels in retinopathy images [42]. The gray value distribution of the blood vessel's cross-section can be regarded as a Gaussian model in the blood vessel contour. The second-order Gaussian derivative can improve the segmentation accuracy of low-quality retinopathy images [43]. The blood vessel contour model can be regarded as a synthesis of blood vessel intersections and branches. More complex models, such as geometric deformation models and parametric deformation models, can also segment blood vessels [44].

2.3.2 Artery/Vein classification

Artery-Vein classification, a usable way of making retinopathy diagnosis in the early stage, distinguishes blood vessels inside the fundus images as two types: arteries and veins. As shown in Figure 6 [45] below, the left is the fundus image, and the right is the corresponding Artery/Vein classification result. There are already some related results, for example, The Arteriolar-to-Venular diameter Ratio (AVR) [46], as a discriminating parameter, whose size could decide various types of illness such as diabetes and cardiovascular diseases. Under typical situations, the DRIVE dataset can be used to perform blood vessel segmentation tasks. However, for the artery-vein classification task, this dataset is no longer suitable. Therefore, more datasets have shown up. The DRIVE-AV dataset [47], with more artery/vein labeling information at pixel level added inside, contains twenty images respectively in the training and testing dataset. For the LES-AV dataset [48], its labeling information is the same as in DRIVE-AV. Both are at pixel levels, but the number of annotated images is 22. The INSIPRE-AVR dataset [49] and the private IOSTAR dataset [50] respectively contain 40 and 24 images. The first one's labeling info is in the centerline level. Two experts annotate the second one's data samples.

Galdran et al. [45] propose that Artery/Vein (A/V) classification is a segmentation problem with four classes to be classified, which are artery, vein, background, and uncertain. Without segmentation to the blood vessels first, they directly choose one network similar to the U-Net to classify arteries and veins. Besides, Raj et al. [51] proposed one network called AV-Net (Artery-Vein Net), using the Res-Net 50 as the principal network used for training. Then using squeeze-and-excitation blocks to learn featural weights by the network's loss, thus making useful features maps' weights large. Finally, differently scaled feature maps are processed by upsampling first, then they are merged to the input image's size for getting the segmentation map. This network has some special needs for inputs, which requires one segmented vasculature map. For the vessel segmentation and A/V classification tasks, they can be performed inside one multitask network simultaneously,



(a) (b)

Figure 6. Artery-Vein classification results: (a) Fundus image (b) Classification results. [45]

proposed by Ma et al. [52].

All in all, A/V classification is a research domain full of visions from all methods mentioned above. More general methods for this task are directly training the A/V classification network. One separate vessel segment with arteries and veins, as a problem of A/V classification, still exists from existed works.

2.4 Active learning approaches

For the retinal image analysis, The use of DNN and enough annotated images can make sure that the training model can get good performance. However, the performance of a DNN is not infinitely ascending with the increase of annotated data, and its performance has corresponding bottlenecks. Getting a large number of annotated samples is infeasible in practice due to the huge annotating cost by experts.

2.4.1 Introduction to active learning approaches

Therefore, supervised learning generates considerable annotation costs. AL is suitable for solving this problem. Burr Settles [7] has introduced active learning in detail, one of the machine learning subareas. In the field of optimal experimental statistics, it is also called query learning. The learning and the selecting components are two necessary and important active learning algorithms modules. From the unlabeled sample set, active learning

uses selection strategies to choose some samples actively, then provides them with experts in related fields to mark. The learning module receives labeled samples from the training dataset to train the learning model. When the learning module meets the termination conditions, the program stops running. Otherwise, the program continuously repeats the above steps for obtaining more labeled samples used for training. AL algorithms have an important assumption: if the learning algorithm is allowed to learn the data it is interested in. It will achieve good results through less training with less amount of data.

As shown in Figure 7 [7], the AL method is an iterative and interactive training process, which is mainly composed of five core parts: unlabeled pool (denoted as U), select queries (denoted as Q), human annotator (denoted as S), labeled training set (denoted as L), and machine learning model (denoted as G). AL combines the above five parts into one process, including updates the model performance, unlabeled sample pool, and labeled data set in a continuous iterative training method. The sequence of AL is shown in Figure 7 [7], which lasts until the target model reaches the preset performance or no longer provides annotation data. Ideally, the number of labeled samples keeps increasing during each iteration, and the model performance also improves. In practical applications, the annotator's accuracy should be ensured as much as possible, aiming to alleviate the model learning error caused by incorrectly labeled samples in the early training stage [7].

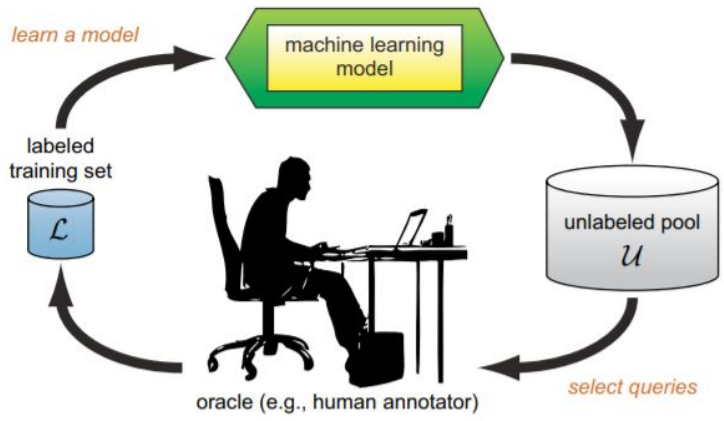


Figure 7. The process of pool-based active learning cycle. [7]

2.4.2 Scenarios of active learning

AL mainly considers the following three scenarios, as shown in Figure 8 [7].

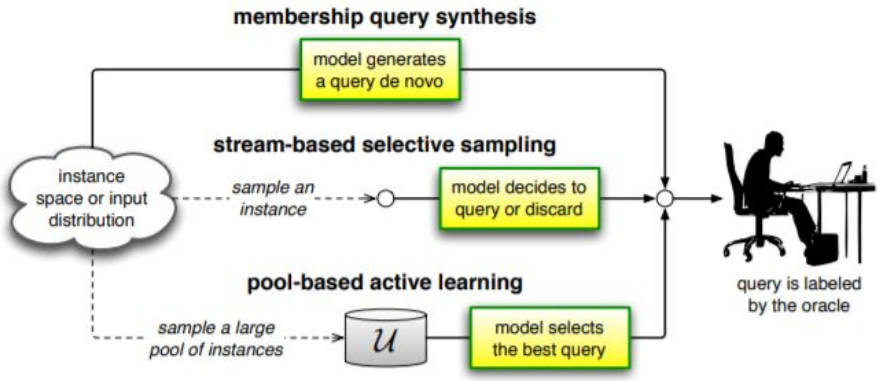


Figure 8. The process of pool-based active learning cycle. [7]

Membership Query Synthesis is the earliest active learning scenario proposed by Augluin et al. [53]. Under this setting, the learner randomly selects specific query instances and then gives them to an oracle for marking. Even the learner itself can generate samples to be labeled. The method achieved good results at the time, but there are certain problems: the query instance experts selected by the learner cannot be identified, and it does not have a fixed semantics [54].

Stream-based Selective Sampling is proposed by Atlas et al. [55]. This scenario supposes many unlabeled instances have already existed. Under this setting, all instances are judged by the learner in turn, and the learner is responsible for judging whether these instances are sent for labeling. The learner usually uses a real-valued function for measurement when making judgments. Instances with higher scores will be sent for marking. This method is also widely used in display, especially in classification algorithms, such as speech tagging and sensor scheduling.

Pool-based sampling [56] usually has a large amount of available unlabeled data. Under this setting, the unlabeled data set can be called a pool, and a real-valued function can still be used to select a query instance each time. This method is most widely used in the real world, such as text recognition, information extraction, and image video classification [57].

2.4.3 Conventional and deep active learning query strategies

The selection strategy of the sample directly determines the degree to which the model can save labeling costs. For example, using the uncertainty strategy can save more an-

notation costs than the random sampling strategy [58, 59]. Because the random sampling strategy neither utilizes the model's prediction information nor many unlabeled sample pools' structure information. Random sampling determines the samples to be labeled first. The uncertainty strategy interacts with the model's prediction information, prioritizing to select samples with the best value for the current model. There are several classic screening strategies:

Random Sampling [58]: this strategy does not need to interact with the model's prediction results. It directly selects samples from the unlabeled sample pool through random numbers, then lets experts label them. It is often regarded as a baseline method in active learning.

Uncertainty Strategy [59]: this strategy assumes that the sample closest to the classification hyperplane, which has a richer amount of information than the classifier. According to the current model's prediction value of the sample, the most uncertain sample is selected. This strategy includes some basic measurement indicators:
1) Least Confidence (LC) uses the additive inverse of the maximum predicted probability as the sample's uncertainty score.
2) Margin Sampling (MS) thinks that the closer the sample is to the classification hyperplane, the higher the uncertainty. It is often combined with SVM and used to solve the binary classification task, but it does not perform well on the multi-classification task.
3) Multi-Class Level Uncertainty (MCLU) is an extension of MS in multi-classification problems. MCLU selects the two samples farthest from the classification interface and uses their distance difference as the criterion.

Query By Committee (QBC): QBC [60] is a sampling strategy based on version space reduction. Version space refers to a collection of a series of different types of benchmark classifiers. The core idea is to select unlabeled samples that can minimize the version space preferentially. QBC includes two basic steps:
1) Using multiple models to form a committee.
2) All committee models sequentially predict unlabeled samples and prioritize screening out the most inconsistent samples for labeling.

Since QBC needs to train several models in the actual application process, it has high computational complexity. To solve this issue, Entropy Query-by-Bagging (EQB) and Adaptive Maximize Disagree (AMD) [61] are proposed for alleviating the problem of computational complexity. EQB introduces the bagging inheritance method and bootstrap sampling; AMD mainly focuses on high-dimensional data, dividing the feature space into

a certain number of subsets and constructing committees.

Nowadays, with the boom of the Internet and the continuous improvement of data collection technology, many fields can obtain a large amount of unlabeled data at a low cost. Deep Learning shows its outstanding performance, mainly reflecting in processing high-dimensional data and automatic feature extraction. As for active learning, its main advantage mainly reflects in effectively reducing annotation cost. Therefore, the combination of active learning and deep learning can complement their advantages to some degree. The Deep Active Learning (DAL) method shows up to solve the labeling cost problem in recent years. Next, the content introduces several query strategy optimization methods in DAL further:

Batch Model Deep Active Learning (BMDAL) [62]. For traditional AL methods, their way of querying samples is through one by one, which leads to the existing model's frequent training and little change to samples of the training dataset. If applying this way to the DL model, the training of this model can be inefficient. Meanwhile, overfitting can happen during the training process. In the BMDAL algorithm process, for each step of acquiring samples, one candidate set $\mathcal{B} = \{x_1, x_2, \ldots, x_b\} \subseteq U$ and one deep model $f_\theta(L)$ is used as inputs of the query function a. The outputs of the query function is a batch of selected samples $\mathcal{B}^* = \{x_1^*, x_2^*, \ldots, x_b^*\}$ used for making annotations. The whole BMDAL process mentioned above can be described by:

$$\mathcal{B}^* = \underset{\mathcal{B} \subseteq U}{\arg\max} a_{\text{batch}} \left(\mathcal{B}, f_\theta(L)\right). \tag{1}$$

Density-based Methods [62].This method takes advantage of one dataset called the core set. The core set selects the most representative samples from the original dataset. In the feature space, these selected samples can represent the original dataset's distribution and reducing the cost of AL annotations.

Uncertainty-based and hybrid query strategies [62]: As a popular query strategy of AL, the approach based on uncertainty is simple to understand and requires low computation complexity. Some machine learning models such as SVM [63] or K-Nearest Neighbor (KNN) [64] can precisely obtain uncertainty like margin sampling, entropy, and least confidence using uncertainty-based methods.

Many DAL [65–68] methods apply uncertainty in the query strategy for sample selection. However, as discussed in the BMDAL, this strategy leads to a deep learning model's insufficient training because it does not take corresponding knowledge related to data dis-

tribution(e.g., diversity) account. Therefore, for taking factors of uncertainty and diversity both into consideration, the hybrid strategy uses these two factors both to query a batch of samples.

Deep Bayesian Active Learning (DBAL) [62]. As is mentioned in the last paragraph, most current classical AL algorithms are based on uncertainty to query valuable samples. To change this situation, deep active learning using Bayesian inferencing shows up. The DBAL can be described by:

$$p(\theta \mid X, Y) = \frac{p(Y \mid X, \theta)p(\theta)}{p(Y \mid X)}. \tag{2}$$

Assuming that there are two parameters: X and Y, X represents the input set, Y represents the output belonging to different classes. With this assumption, $f(\mathrm{x}; \theta)$ can express the neural network based on probability. Besides, $p(\theta)$ represents the prior of the Bayesian model, generally existing inside the Gaussian parameter space θ, $\mathrm{softmax}(f(\mathrm{x}; \theta))$ determines the likelihood $p(\mathrm{y} = c \mid \mathrm{x}, \theta)$. The purpose of DBAL is to gain the posterior distribution related to θ. For a given data point $\mathrm{x}^*, \hat{\mathrm{y}}$, whose posterior distribution can be predicted by:

$$p\left(\hat{\mathrm{y}} \mid \mathrm{x}^*, X, Y\right) = \int p(\hat{\mathrm{y}} \mid \mathrm{x}, \theta)p(\theta \mid X, Y)d\theta = \mathbb{E}_{\theta \sim p(\theta \mid X, Y)}[f(\mathrm{x}; \theta)]. \tag{3}$$

Other mainstream active learning methods. Huang et al. [69] propose an active learning method for deep neural networks, which can transfer the trained deep model to different tasks with fewer samples, thereby reducing the learning of deep neural networks cost. Huang et al. [70] propose a method that combines active learning and matrix completion technique, which can effectively use the label information when severe feature loss has been suffered, thus saving the cost of feature extraction. Chu et al. [71] believe that active learning strategies applied to different data sets have effective experiences. These experiences can be transferred to other data sets to improve the performance of models or strategies. The authors try to migrate the model to different data sets. The experiment proves that most current strategies have effective experience, that experience can also be transferred to different data sets, and improve feature learning tasks' performance.

Neural Architecture Search (NAS)+Deep Active Learning. In all the active learning methods mentioned above, the task model is selected from existing models based on prior knowledge, and the network structure of the model is fixed. There are the following shortcomings:

- There are no ready-made models available in many fields, such as the medical image field.
- In the early iteration process, the number of labeled samples is small, and the fixed network structure model may fall into overfitting.

NAS [72] can effectively solve the shortcomings caused by the fixed network structure. In the case of uncertain network and structure, a recurrent network is used as the controller's field to generate the network structure to construct the sub-neural network. The accuracy rate after training the sub-network is used as the controller's feedback signal, and the controller is updated by calculating the gradient of the strategy, thus continuously iterating the loop.In the next iteration, the controller will have a higher probability of proposing a high-accuracy network structure. In short, as time goes by, the controller will continuously learn to improve search results.

Geifman et al. [73] first tried to apply NAS to the active learning method. Thus the network structure of the model can adapt to the newly added annotation data. The experimental results indicate that the active learning method's efficiency has been improved, especially after joining NAS to the fixed network structure's active learning method.

2.5 Active learning applications on the retinal image analysis

Some related works are combining active learning with retinal image analysis together in recent years. These works mainly solve three retinal image analysis tasks: retinal image classification, retinal image segmentation, and retinopathy prediction.

Ayerdi et al. [74] propose one active learning method based on uncertainty to solve the blood vessel segmentation task. The random forest(RF) classifier is the model trained during the whole learning process. As the input of this classifier, features are extracted by simple statistical methods and undirected morphological operators. These operators are mainly gained by computing the green component of the image.

Mahapatra et al. [75] propose one active learning(AL) network to train one deep learning model, thus solving medical image classification and segmentation tasks. They use the conditional generative adversarial networks Conditional Generative Adversarial Networks (CGANs) to generate samples with various features, then use the Bayesian neural network to select the most informative samples used for training. This method can achieve good performance by only 35% of the whole lung X-ray image dataset.

To realize the retinal image's retrieval and automatic annotation, Punithavathi et al. [76] propose one method combining Support Vector Machine(SVM) and Active Learning(AL), making annotations to the retinal image automatically. They also use Bray Curtis distance as the similarity measurement to retrieve retinal images.

Hemelings et al. [77] try to combine transfer learning with active learning to solve the glaucoma diagnosis task. They use the ResNet-50 as the Benchtrack model, training it and making it capable of classifying two sorts of fundus images: the glaucomatous ones and non-glaucomatous ones. The active learning strategy uses uncertainty as the standard to select samples. Selected samples generally are heat maps generated by the deep learning classifier, which can help experts assess their decisions.

Li et al. [78] propose one improved U-NET model. This model can accelerate the calculation speed of getting blood vessel segmentation results by modifying the original U-NET structure. Then they still adopt the uncertainty-based approach as the query strategy of active learning. Experimental results indicate that this algorithm exceeds supervised learning on the accuracy of blood vessel segmentation.

Some new query strategies have been proposed in recent years, for example, Ozdemir et al. [79] propose one query strategy combining representativeness with uncertainty, aiming for finding the most valuable samples to be annotated. Their representativeness measurement is mainly based on Bayesian sampling. It has used autoencoders with the function of maximizing information. Based on this improvement, their algorithm shows better performance on medical image segmentation, especially when compared to traditional representativeness measurement methods.

2.6 Summary

This chapter mainly discusses contents around two aspects: retinal image analysis and active learning. Related to the retinal image analysis methods, contents primarily focus on retinal blood vessel segmentation and classification, discussing conventional image processing performance and CNN methods on these two tasks. The method comparisons show that CNNs generally are better than traditional methods on retinal image segmentation and classification tasks. Next, scenarios, query strategies of active learning are discussed. Finally, making comparisons upon the conventional and the deep active learning methods, results indicate that a combination of active and deep learning has better performance on the retinal image analysis task.

3 RETINAL BLOOD VESSEL SEGMENTATION WITH ACTIVE LEARNING STRATEGIES

The study mainly adopts a U-net [80] architecture network as the model to accomplish the blood vessel segmentation task. In the first place, model training starts using few annotated data to get an initialized model with the essential accuracy, then selecting the most valuable samples for experts to make annotations. The implemented active learning strategies are based on uncertainty in blood vessel segmentation results. Samples annotated by experts rejoin the original training samples to build one new training dataset, and then the new dataset does fine-tuning to the existed model. All steps mentioned above are an iterative process, which terminates when there are no unannotated samples anymore. The following contents will introduce several aspects involved in the proposed algorithm: data preprocessing and augmentation, U-net [80] architecture segmentation network, pre-trained model, continuously fine-tuning, and active learning sampling strategies.

3.1 General framework of the proposed method

The main framework of the proposed method is displayed in Figure 8

For the blood vessel segmentation task, the DRIVE dataset can be used as an example. This dataset gets 20 images included used for training, and another 20 images included used for testing. The steps of the proposed method are:

Step1: Randomly selecting N fundus images from the DRIVE training dataset, these images and their corresponding ground truths are used as the annotated dataset used to train the pre-trained model.

Step2: Performing image preprocessing for all fundus images in the annotated dataset, preprocessing ways generally includes gray-scale transform, CLAHE, and other methods, then applying the data augmentation to preprocessed images and their ground truths by randomly cropping the full images into small patches, finally sending them inside the network for training.

Step3: For the DRIVE dataset, using the remaining 20-N fundus images as the unannotated dataset, still doing the same image processing operations as Step 2 does to each unlabeled fundus image, then predicting all unlabeled samples to get valuable informa-

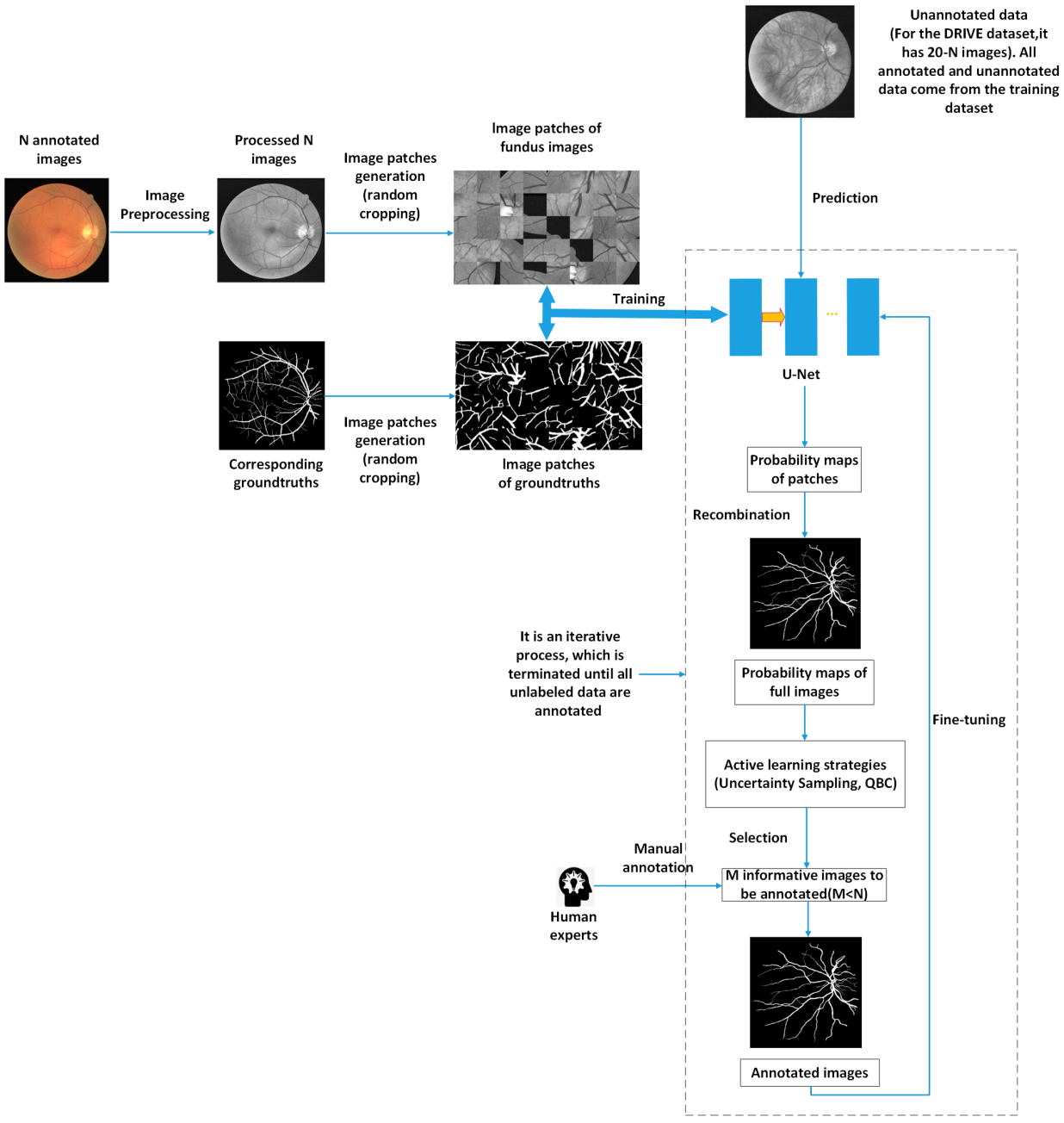


Figure 9. The diagram of the proposed method.

tion like probability maps for sampling, then adopting active learning strategies like uncertainty to identify which M (M<N) samples are about to be annotated next. Here the algorithm uses the ground truths of unlabeled data as the human expert annotation results. Finally, sending the annotated images and their annotated labels back to do the fine-tuning for the pre-trained model. Step3 is an iterative process, which comes to an end when all unlabeled data are annotated.

Everything introduced above is the main flow of the blood vessel segmentation algorithm using active learning. The following chapters introduce the implementation in detail.

3.2 Data preprocessing and data augmentation

3.2.1 Data preprocessing

The first step of data preprocessing is to resize all images from datasets into a unified format. For the convenience of input to the training network, some input pictures from datasets are resized into equal width and height. Other input images with a larger resolution are reduced to a specific size. Meanwhile, the aspect ratio of all input images remains unchanged.

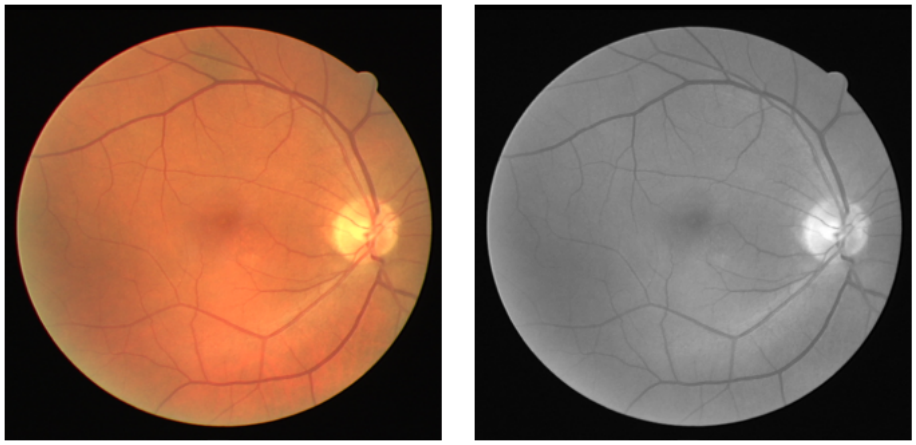
Next, the algorithm uses four ways to preprocess the original images: gray-scale transformation, normalization, Contrast Limited Adaptive Histogram Equalization (CLAHE), and gamma transformation.

For the gray-scale transformation of the original images, the implementation uses the simple addition of the R, G, and B components from images. The weighted sum of these three components has specific coefficients. The way of making additions is:

$$I = 0.299 \times R + 0.587 \times G + 0.114 \times B \tag{4}$$

where R, G, B are values of three components from the RGB color space. I offers the intensity information by adding values of R, G, B three channels. The coefficients 0.299, 0.587, and 0.114 of this formula are acquired by the human eye's different perceptions of color. By performing the same addition calculation to each pixel of one image, the intensity values of all pixels can combine into a gray-scale image.

The role of gray-scale transformation is to convert RGB images into gray-scale images. Meanwhile, the transformation also eliminates image saturation information while retaining brightness. The contrast effect of the gray-scale transformation on the fundus image is shown in Figure 10.



(a) (b)

Figure 10. Gray transformation results: (a) Fundus image (b) Transformation results.

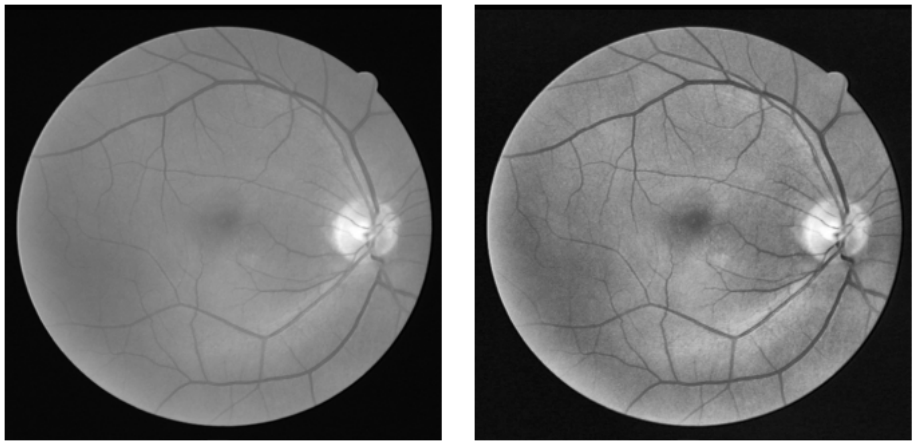
The normalization here generally refers to the Min-Max Normalization, also known as dispersion standardization. It is a linear change to the original data, thereby making the values of the result mapped between 0 and 1. The mathematical expression of the Min-Max Normalization is:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{5}$$

where x' is the normalized value, x is the original value, and generally, the x is a sequence of data.

The Contrast Limited Adaptive Histogram Equalization (CLAHE) originates from Histogram Equalization (HE), which is performed by using the gray-scale distribution map to identify one projection line for gray-scale transform, aiming for improving the image contrast. The projection line here refers to the Cumulative Distribution Function (CDF) histogram. However, the HE is the global contrast adjustment method, which cannot efficiently improve the local contrast. The Adaptive Histogram Equalization (AHE) aims to improve the local contrast of the image by dividing the whole picture into few small patches, then doing HE processing to each patch respectively. A disadvantage of AHE

is that its improvement for local contrast is too significant, which will contribute to the distortion of images. Therefore, CLAHE appears, the difference between CLAHE and another two algorithms is that the former imposes the restriction on the local contrast, and it adopts interpolation to accelerate the calculation. It can efficiently enhance or modify the local image contrast to acquiring more related info about edges for better segmentation results. Meanwhile, CLAHE can remedy the problem of amplifying in AHE algorithm. After implementing the CLAHE algorithm to the gray-scale fundus image, the processed results are shown in Figure 11.



(a) (b)

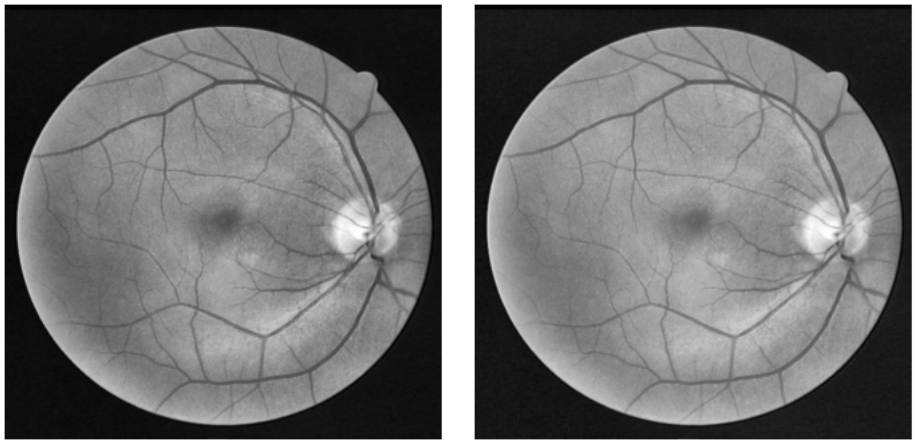
Figure 11. CLAHE transformation results: (a) Gray-scale fundus image (b) Transformation results.

Gamma correction performs nonlinear operations to gray-scale values of input images, making the exponential relationship between gray-scale values of input images and output images. The corresponding exponential relationship can be described by:

$$V_{\text{out}} = AV_{\text{in}}^{\gamma} \tag{6}$$

where the γ is the index, V_{in} and V_{out} represent output image gray-scale value and input image gray-scale value. A is a constant, and it usually takes 1. The primary function of gamma correction is to perform image enhancement, improving the quality of details in dark parts of the image. In short, through the nonlinear shift, the linear response of the image from exposure intensity becomes closer to the one perceived by the human eye. That is, this correction will correct the bleached (camera exposure) or too dark(underexposed) images. The comparison of the before and after effects of gamma correction on images is

shown in Figure 12.



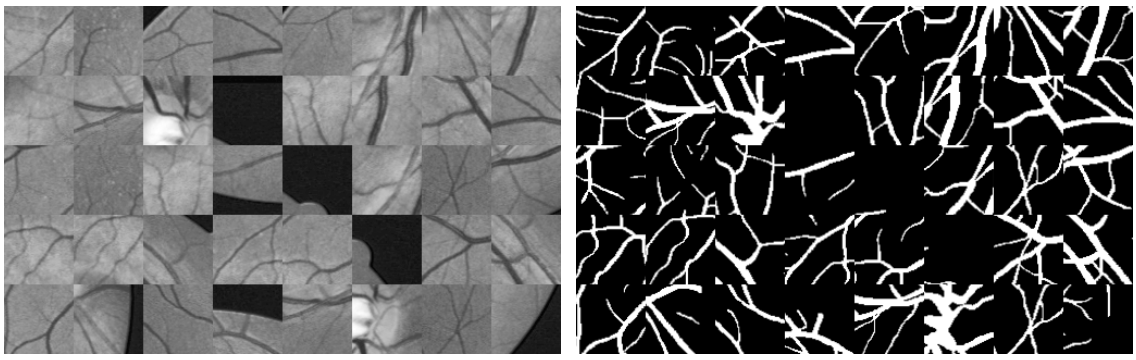
(a) (b)

Figure 12. GAMMA transformation results: (a) Fundus image processed by CLAHE (b) Transformation results.

3.2.2 Data augmentation

For data augmentation, the fundus datasets used for training CNN model are small, generally only having dozens of pictures. The scale cannot satisfy the needs of the CNN model requiring a vast amount of training samples. Therefore, it is necessary to expand the original datasets by generating more training samples. Here, the data augmentation aims to create more image patches of fundus images by random cropping, then feed them to the CNN model for training.

The current project randomly extracts N patches for each fundus image to build the baseline classifier. The N here should be greater than or equal to the number of splits, which is the number of patches obtained by averaging the original image using a fixed patch size. These N patches can be called Region of Interest (ROI). Their extraction adopts the rectangle to indicate the ROI area, which can be described by a quaternion: [x, y, width, height]. (x, y) here refers to the left upper coordinates of the ROI rectangle. For the current project, its value can be any pixel coordinate of the whole fundus image. Width and height are attributes information of the ROI rectangle. For the DRIVE dataset, 40 randomly selected training patches and their corresponding ground truths are shown in Figure 13.



(a) (b)

Figure 13. Training samples from one fundus image: (a) Gray-scale image patches. (b) Corresponding groundtruths.

3.3 U-net architecture segmentation network

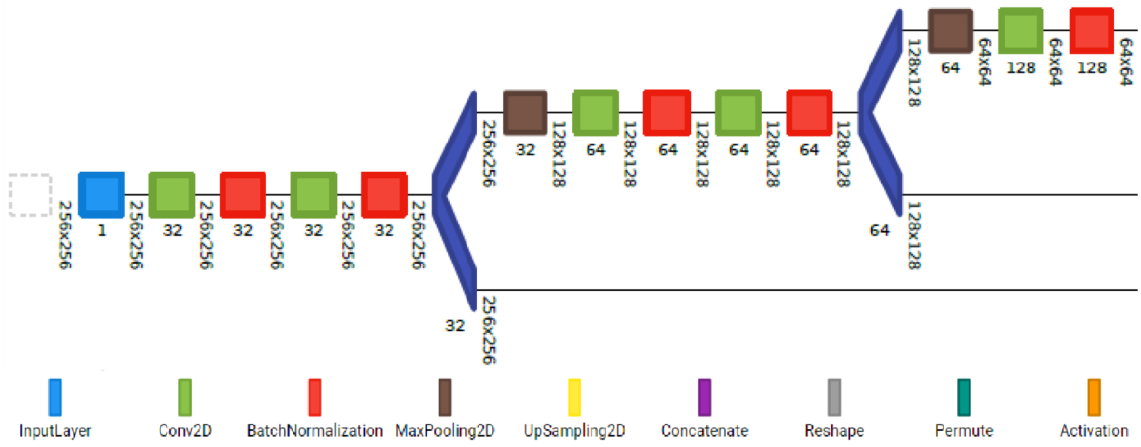
For the blood vessel segmentation task, its implementation relies on a U-net architecture like network, the architecture of which is shown in Figure 14.

The current network used in the study is shown in Figure 14. The digits between layers represent the size of feature maps. For numbers under blocks, they represent the number of filters of each layer. Compared with the U-net, the current network still adopts the encoder-decoder structure, having two downsampling layers and two upsampling layers. The left two layers are called the encoder, responsible for extracting simple convolutional features from input data. The other two layers are called the decoder, which resumes the image back to its original size by transposed convolution. Meanwhile, the acquired feature maps from the encoder will be concatenated to their corresponding upsampling layers. Therefore, the working process of this network is to use the encoder to perform the feature extraction first, then utilize the decoder to resume the extracted feature maps to their original sizes, finally proceed with the classification for each pixel of the input image.

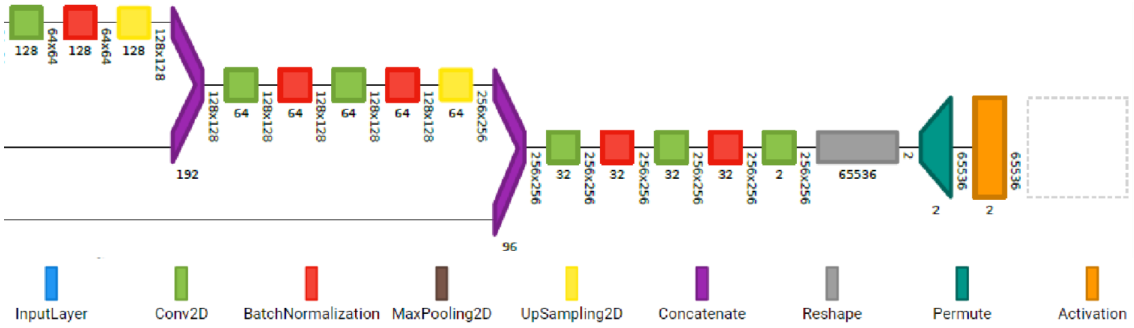
The current network shown in Figure 14 has two differences from the original U-net. The first one is that multiple BatchNormalization layers are added to the network. The goodness of the BatchNormalization has five points [81]:

- The convergence of the training network significantly speeds up.
- The network performance is improved due to the prevention of the overfit.

- The adjustment of training parameters, such as learning rate, is simplified. The training process of this network can apply a significant learning rate.
- The problem of vanishing gradients can be prevented by using the BatchNormalization.



(a)



(b)

Figure 14. The visual structure of U-net architecture segmentation network. (a) The left part of the network. (b) The right part of the network.

Moreover, the output segmentation map of the modified network is also different from the original U-net. For example, if the input size of the image is M×N×1 (width×height×channel), the size of the output segmentation map will be (1, M×N, 2). The output segmentation map is the input image's M×N pixels prediction probabilities. If it involves the blood vessel segmentation task, every pixel of this image has two prediction probabilities (about labels 0 and 1).

For the blood vessel segmentation task, the loss function used here is categorical crossentropy. It is principally used for evaluating the difference of the probabilistic distributions acquired from training and ground-truth data. It describes the distance between the actual

output(probabilities) and the excepted output(probabilities), which means if the cross-entropy is much smaller, the two probabilistic distributions are closer. The categorical crossentropy is decided by:

$$loss = -\sum_{i=1}^{n} \hat{y}_{i1} \log y_{i1} + \hat{y}_{i2} \log y_{i2} + \cdots + \hat{y}_{im} \log y_{im} \tag{7}$$

where n is the number of samples, m is the number of classes. $\hat{y}_{i1}, \hat{y}_{i2}, ...\hat{y}_{i2}$ is the prediction values(generally is probabilities), while $y_{i1}, y_{i2}, ...y_{i2}$ are corresponding ground-truth labels of $\hat{y}_{i1}, \hat{y}_{i2}, ...\hat{y}_{i2}$.

3.4 Transfer learning + continuously fine-tuning

As introduced in Chapter 2, active learning generally has few annotated data used for training. To fully utilize annotated data and guarantee the training model accuracy on the new task (training on the unannotated data), a combination of transfer learning and fine-tuning is used as the training strategy of the current project. The combination is involved with the transferability of each layer in deep learning, which is discussed by Bengio et al. in [82]. They performed the experiments as Figure 15 shows, using the ImageNet dataset to train one CNN model in four different ways, thus realizing the image classification task.

There are four CNN models acquired after the training process:

- The basic model baseA in domain A.
- The basic model baseB in domain B.
- In domain B, utilizing baseB parameter initialization on the first n layers of the training model, making the remaining layers untrainable and doing fine-tuning to them.
- In domain B, utilizing baseA parameter initialization on the first n layers of the training model, making the remaining layers untrainable and doing fine-tuning to them.

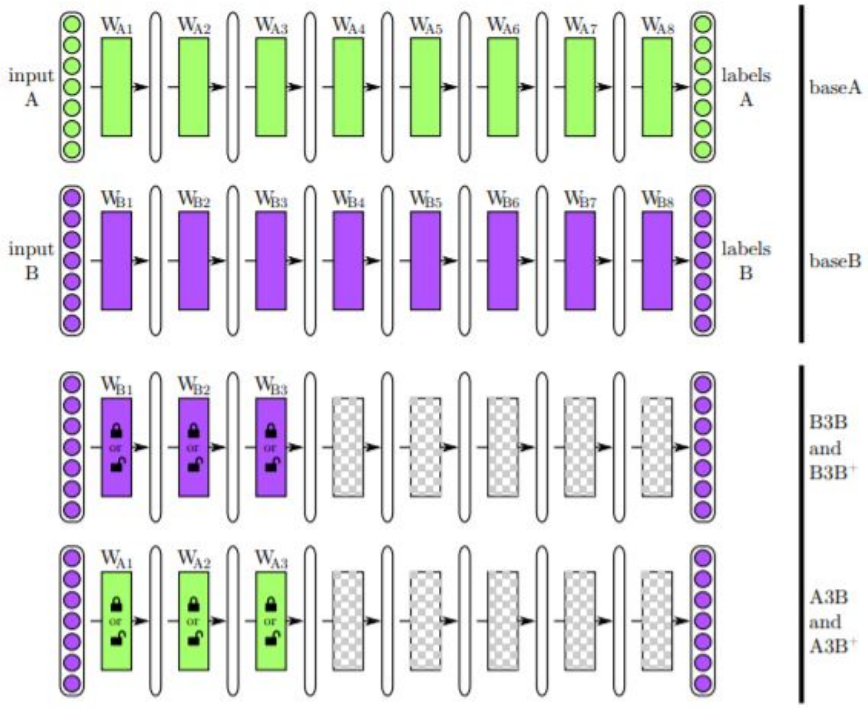
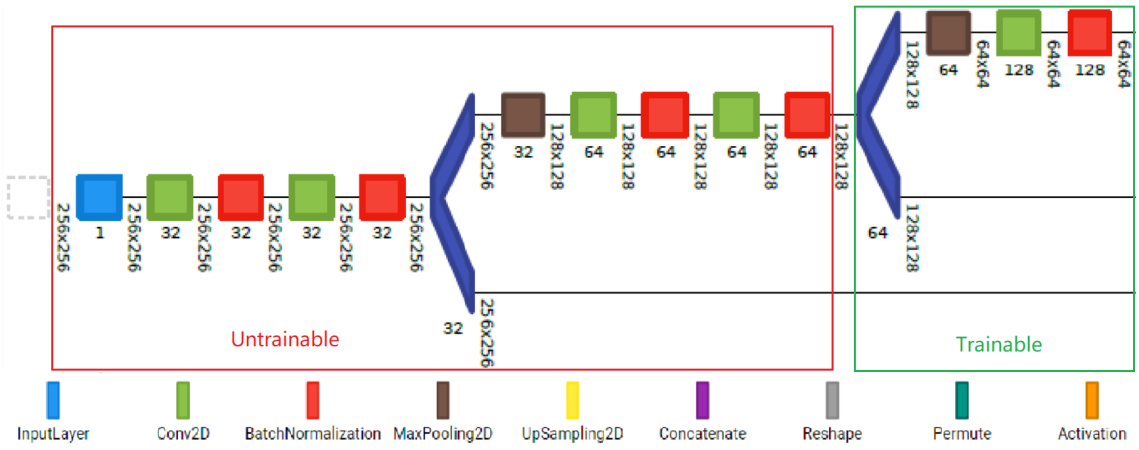


Figure 15. Four of the experimental treatments and controls. [82]

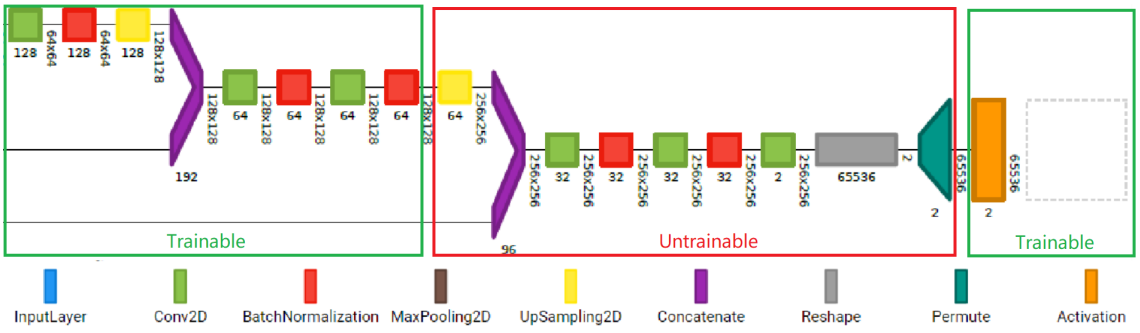
The domain here comes from transfer learning. Transfer learning has two critical concepts: domain and task. The first one refers to a specific field at a particular moment. For example, interviews on books and TV series are two different domains. The task here refers to the thing about to do. For instance, emotion analysis and instance recognition are two various tasks. Therefore, the experiment presented above has two domains(A and B) and one task(image classification). According to the conclusion drawn by the authors, the extracted features from the first layer in the CNN model has little relationship with the image dataset. In contrast, the network's last layer is closely related to the selected dataset and task goals. [82] respectively call features from the first and the final layer as the general features and the specific features.

Therefore, according to the point of view in [82], the combination of pre-training and fine-tuning can also be applied in the U-net architecture segmentation network. There are few steps for its implementation. Firstly, using annotated data to train the model so that it has essential accuracy. Then, freezing corresponding layers as described above, which include two parts. The first part is in front of the second downsampling layer, responsible for general feature extraction. The second part is between the last upsampling layer and the final activation layer, responsible for transforming features acquired from the final feature fusion. Except for these two parts, other parts will not be frozen, which are in charge of specific feature extraction (generally refers to deeper semantic features)

and feature fusion. Making layers frozen here means making layers untrainable. Trainable and untrainable components are shown in Figure 16. After making corresponding layers untrainable, the next step is to select images for annotation by using active learning strategies. Then the chosen images are divided into image patches and sent back to the network for fine-tuning. The fine-tuning in the current project is a cyclic process, which is terminated until all unlabeled samples are annotated.



(a)



(b)

Figure 16. Trainable and untrainable parts of the segmentation network. (a) The left part of the network. (b) The right part of the network.

3.5 Sampling strategies

The current project mainly uses four active learning sampling strategies. The first three strategies are based on uncertainty, and the last strategy is a method based on density weighting, which considers noise samples while considering uncertainty.

Uncertainty Sampling (US) is one of the most general sample querying strategies. The

main goal is to return the most confusing or the most informative samples to experts, thereby gaining the maximum gain. US generally includes three approaches: least confidence, margin sampling, and information entropy.

For the least confidence method, its uncertainty confidence score is decided by:

$$x^*_{\mathrm{LC}} = \mathrm{argmax}_x \, 1 - P_\theta(\hat{y} \mid x), \quad \hat{y} = \mathrm{argmax}_y \, P_\theta(y \mid x) \tag{8}$$

where x is the unlabeled sample, $\hat{y}$ is the category with the x's highest prediction probability coming from the training model, and x_{LC} represents the most uncertain samples for the current training model. Therefore, this formula is used for identifying the instances which have the minimum probability for their most confident category.

Even though the least confidence is straightforward, it only considers samples with the highest model prediction probability but low confidence. For those samples with lower prediction probability, they are not within the scope of consideration. Therefore, margin sampling shows up, and it is given by:

$$x^*_{\mathrm{M}} = \mathrm{argmin}_x \, P_\theta \left(\hat{y}_1 \mid x\right) - P_\theta \left(\hat{y}_2 \mid x\right) \tag{9}$$

where x is the unannotated sample, $\hat{y}_1$, $\hat{y}_2$ are the first and second most probable class labels of it. x^*_M is the most uncertain sample, which is decided by the difference between the first and second highest probability of x.

Maximum entropy can be another measurement method of uncertainty. It considers one sample's prediction probabilities of all categories. The entropy can describe the confusion degree of one system. When the probabilities are the same, the entropy takes the maximum value. When all probabilities are concentrated, the entropy takes a smaller value. Moreover, when the predicted probabilities are similar, the prediction of the current sample is useless, and its uncertainty is very high. The mathematical expression of maximum entropy is given by:

$$x^*_{\mathrm{H}} = \underset{x}{\mathrm{argmax}} - \sum_i P_\theta \left(y_i \mid x\right) \log P_\theta \left(y_i \mid x\right) \tag{10}$$

where x is the unannotated sample, $P_\theta\left(y_i \mid x\right)$ is all prediction probabilities for sample x. This method considers all prediction probabilities for one sample, using entropy to describe the degree of confusion. The higher entropy brings this system with more even probability distribution and higher uncertainty.

The last three strategies are to select samples with significant uncertainties to improve the performance of the segmentation network. However, sometimes, samples with the highest uncertainties may also be noisy samples, and they cannot improve the model's performance. Therefore, the active learning strategy can consider the overall distribution of the samples more.

The density-weighted method considers samples with the highest uncertainty and the highest representativeness at the same time. Its mathematical expression can be described by:

$$x_{\mathrm{ID}}^{*}=\arg\max_{x}\Phi_{A}(x)\times\left(\frac{1}{U}\sum_{u=1}^{U}\operatorname{sim}\left(x,x^{(u)}\right)\right)^{\beta} \tag{11}$$

where $\Phi_A(x)$ represents the information volume identified by some basic sampling strategies(US, QBC), $\frac{1}{U}\sum_{u=1}^{U}\operatorname{sim}\left(x,x^{(u)}\right)$ represents the average similarity between x and each unlabeled sample $x^{(u)}$ in the set U. $\operatorname{sim}(*)$ is the similarity calculation function, the higher value it takes, the more similar it will be. β is the control parameter. Here $\operatorname{sim}(*)$ can be defined by using Euclidean distance.

One thing that needs to be noted here is that the density-weighted strategy aims to select samples with high uncertainty, which are also very similar to most unlabeled samples. Therefore, two primary components: information volume $\Phi_A(x)$ and similarity $\operatorname{sim}(*)$, need to take the highest values. However, the higher value it takes for the Euclidean distance, the less similar it will be. Some transform can be done to the Euclidean distance, which is given by:

$$\operatorname{sim}\left(u_i,u_j\right)=\frac{1}{1+\mathrm{d}\left(u_iu_j\right)}\in(0,1]. \tag{12}$$

This formula normalizes the acquired Euclidean distance $\mathrm{d}\left(u_iu_j\right)$, making the range of similarity restricted within (0, 1]. Compared with the range of Euclidean distance [0, inf], the normalized result can better quantify the similarity value between vectors. At

this time, the smaller Euclidean distance it takes, the more similarity it will be, and the similarity between two samples u_i, u_j will be higher.

Active learning strategies in the segmentation task mainly utilize information from probability maps. Just like mentioned above, one fundus image has two prediction probabilities for each pixel (for categories 0 and 1). Therefore, one fundus image prediction results include two probability maps: probability map of category 0 and probability map of category 1. To get the uncertainty or other confidence of fundus images, the four strategies mentioned above use these two probability maps to do the calculation. Uncertainty or other confidence of one fundus image is one matrix. For convenience, the average of this matrix is directly taken as the confidence score of the whole fundus image.

4 EXPERIMENTS

This chapter introduces all details about the experiments of active learning strategies, including the used datasets, evaluation criteria used to evaluate experimental results, specific ways of carrying out experiments, and analysis for experimental results.

4.1 Data

Three fundus image datasets are used in the experiments of blood vessel segmentation, and they are DRIVE, HRF, CHASEDB1. More information about these datasets is shown in Table 2. Randomly drawing samples from three datasets mentioned above, fundus images used for the blood vessel segmentation task are shown in Figure 17.

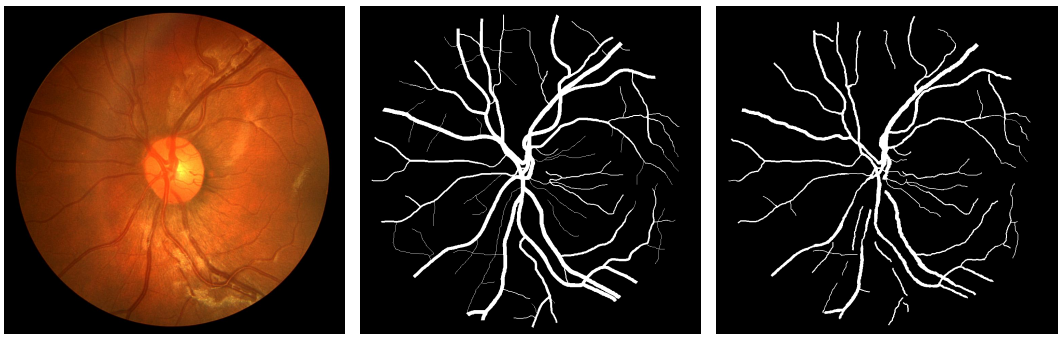
Table 2. Specific details of the three fundus databases. Resolution-resolution of fundus images, the format is width×height, N_{fundus}-the number of fundus images in each dataset, N_{label}-the number of annotated segmentation results in each dataset, N_{experts}-the number of experts making annotations for each dataset.

Database	Resolution	N_{fundus}	N_{label}	N_{experts}
DRIVE	565×584	40	40	1
HRF	3504×2336	25	45	1
CHASEDB1	999×960	28	56	2

One sample fundus image and its corresponding labels are displayed in Figure 18. As seen on the right side of Figure 18, the ground truth is two binary images that only contain two kinds of pixels: white pixels and black pixels. White pixels represent blood vessels in the retina, and their pixel value is 1. Black pixels represent all background areas except for blood vessels in fundus images, and their pixel value is 0.

(a) (b) (c)

Figure 17. Example images from different retinal databases: (a) DRIVE (b) CHASEDB1 (c) HRF.



(a) (b) (c)

Figure 18. One sample image from the CHASEDB1 dataset: (a) True color fundus image. (b) Annotation made by the first expert. (c) Annotation made by the second expert.

4.2 Evaluation criteria

The following indices are used to evaluate the algorithm's performance for evaluation: Specificity, Sensitivity, Pixel Accuracy, and F1-score [83]. More details are about to be introduced in the following contents.

Blood vessel segmentation is a binary classification task. For each pixel, one prediction label(1 or 0) is given as the prediction result, indicating that this pixel is the foreground(blood vessel) or background. Therefore, the segmentation results for the entire image generate two sorts of pixels. The categories are 1 and 0, which can be used as positive and negative categories, respectively. The actual classification results have four types, which are displayed in Table 3.

Table 3. Prediction results for binary classification.

		Predicted		
Actual		1	0	Total
	1	True Positive(TP)	False Negative(FN)	Actual Positive(TP+FN)
	0	False Positive(FP)	True Negative(TN)	Actual Negative(FP+TN)
Total		Predicted Positive(TP+FP)	Predicted Negative(TP+TN)	TP+FP+FN+TN

All evaluation indicators mentioned above can be drawn from this table, which can evaluate the segmentation performance, and the following contents will introduce these indicators one by one.

True Positive Rate (TPR) describes the proportion of the identified positive samples to all positive samples. It is calculated by:

$$TPR = \frac{TP}{TP + FN} \tag{13}$$

where TP and FN represent the number of pixels belonging to the True Positive and False Negative cases.

True Negative Rate (TNR) represents the proportion of identified negative samples to all negative samples. Its calculation formula is:

$$TNR = \frac{TN}{TN + FP} \tag{14}$$

where TN and FP represent the number of pixels belonging to the True Negative and False Positive cases. Moreover, TPR is also called sensitivity, and TNR is called specificity.

Pixel Accuracy (PA) is to calculate the ratio between the number of correctly classified pixels and the number of total pixels. Its calculation formula is:

$$PA = \frac{\sum_{i=0}^{k} p_{ii}}{\sum_{i=0}^{k} \sum_{j=0}^{k} p_{ij}} \tag{15}$$

where k is the number of categories, which is $k + 1$ if getting background included, p_{ii}

represents the pixels that are correctly classified. p_{ij} represents the total number of pixels whose real pixel category is i is predicted to be category j.

F1-score is one measurement indicator of the classification performance. It is the harmonic mean of precision and recall, the maximum is 1, and the minimum is 0. Its value can be acquired by:

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{16}$$

where $precision$ and $recall$ exactly are TNR(specificity) and TPR(sensitivity) mentioned above.

4.3 Description of experiments

Multiple active learning strategies are implemented in this project. To check whether the segmentation network performance is improved after applying these strategies. The experiment uses the random sampling strategies as the baseline, designing different annotation situations in each retinal image dataset. The goal is to make comparisons among various active sampling strategies under diverse annotation circumstances.

For the blood vessel segmentation task, the fundus images are adjusted to the specified size first, and then they are divided into small image patches and sent to the network for training. One thing that needs to note is that the image patch size is critically important, which decides the receptive field of the training network. Too large or too small patches will affect the segmentation accuracy acquired from the training model. The settings of related experimental parameters in this process are shown in Table 4.

The parameter selection in Table 4 has several reasons. For resized resolution, its parameter selection depends on the original resolution's width and height. If they are getting close to each other, the original image can be resized into a square one (DRIVE, CHASEDB1). Otherwise, the aspect ratio remains unchanged when resizing the original image (HRF). Besides, the high-resolution image means high dimensionality of the data, which costs more training time and slows down the convergence speed. Therefore, the high-resolution image can be resized into the image with the aspect ratio remained and not losing distortion. The patch size selection depends on the receptive field mentioned above, the receptive field for blood vessels should have the proper size in human eyes,

which cannot be too large or too small for the training model.

Table 4. The adjustment for the original fundus image size. Original Resolution-fundus image resolution before size adjustments. Resized Resolution-fundus image resolution after size adjustments. Patch Size-the size of image samples sent to the network for training.

Database	Original Resolution	Resized Resolution	Patch Size
DRIVE	565×584	576×576	64×64
HRF	3504×2336	1440×960	96×96
CHASEDB1	999×960	960×960	96×96

During the training process, the whole training dataset is split into two parts: the annotated and unannotated parts. Images from the annotated part are used for training the pre-trained model. The specific task is to specify the specific number of patches generated from each image, then taking advantage of random cropping to get patches from each image. Finally, all acquired patches are used for training the pre-trained model. The initial small learning rate is to make sure the training model incline to convergence. In comparison, the learning rate for fine-tuning is lower than the initial learning rate. The weights acquired from the pre-trained model are better than ones acquired randomly-initialized model, so the lower learning rate can make sure the weights are not changed too much. Therefore, the learning rate for generating the pre-trained model is 0.0005, and the learning rate for continuously fine-tuning is 0.0001. The optimizer is the adam, and the loss function is the categorical crossentropy. Moreover, the number of epochs for both training pre-trained model and doing fine-tuning is 10. Related information such as the division of datasets, various annotated situations is shown in Table 5.

Table 5. The related parameters settings in the training process.N_{sample}-the number of fundus image samples in each dataset. N_{train}-the number of training image samples in each dataset. N_{test}-the number of testing image samples in each dataset.

Database	N_{sample}	N_{train}	N_{test}	The data range of the annotated image
DRIVE	40	20	20	[2, 6, 10, 14, 18]
HRF	45	22	23	[2, 6, 10, 14, 18, 20]
CHASEDB1	28	14	14	[2, 5, 8, 11]

As it can be seen from Table 5, each dataset is divided into training and testing datasets. All annotated samples are taken from the training datasets, and the annotation can be categorized into various situations, including different numbers of annotated images. Multiple models are acquired from annotation situations. Then these models can be used to segment testing samples, thus receiving prediction results to get the values of evalua-

tion criteria. Finally, these values are used for making the comparison between multiple training-generated models.

The design of experiments mainly includes two parts. The first part is to observe the pixel accuracy change with the increase of patches used for training. For each dataset, the specific task is to calculate the pixel accuracy during the training process, and the test dataset is used for acquiring pixel accuracy. All accuracies are recorded only at two stages: after pre-training and when the fine-tuning process is over. The second part is to perform one comprehensive evaluation after the training process. There are many models generated under various annotation situations. Then the research uses them to make predictions for the testing dataset of each retinal image set. Finally, making comparisons over prediction results, its goal is to observe the effects of increasing the number of training images on the active learning strategies.

4.4 Experimental results

This section presents the experimental results on three datasets (DRIVE, HRF, CHASEDB1) for blood vessel segmentation. More specifically, after implementing various active learning strategies, the section will compare their performance of blood vessel segmentation.

The following contents show the experimental results of the second designed part mentioned in Section 4.3: the change of pixel accuracy with the increasing number of training images on the active learning strategies. After setting various annotation situations as shown in Table 5, the performance of multiple active learning strategies on testing accuracy is displayed in Figure 19.

In the subfigure (a) of Figure 19 (DRIVE dataset), when the number of labeled data is 6, 10, 14, or 18, the pixel accuracy of all active learning strategies exceeds or is equal to the benchmark random sampling. For the strategies density-weighted and least-confidence, their pixel accuracy is lower than the random sampling when the number of labeled data is 2. However, with the increasing of labeled data, the pixel accuracy of these two strategies gradually increases and exceeds the random sampling. For the strategy margin sampling, its pixel accuracy is generally upward and exceeds the benchmark (random sampling). However, except for when the number of labeled data is 2, in most cases (when the number of labeled data is 6, 10, 14, or 18), the accuracy of margin sampling is lower than the least confidence and density-weighted. For the strategy entropy sampling, the pixel accuracy is generally upward. Its accuracy only exceeds other strategies in some cases (when the

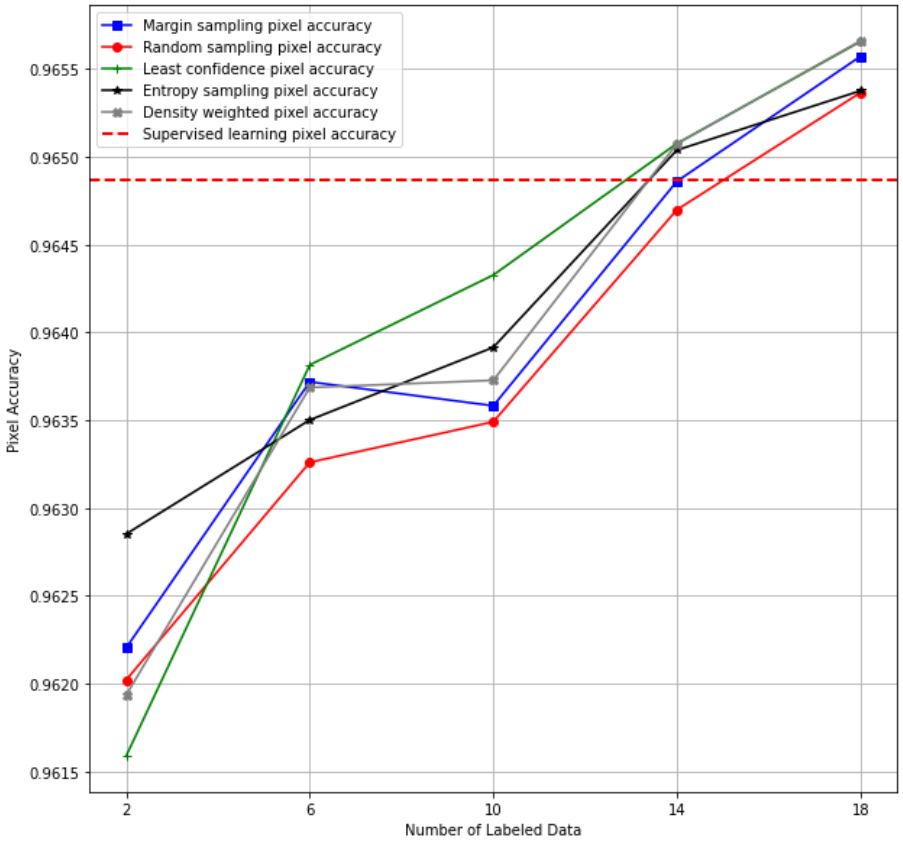
number of labeled data is 2, 10, or 14). In short, the least confidence shows the best performance than other strategies in all annotation cases.

In the subfigure (b) of Figure 19 (HRF dataset), the strategy entropy sampling performs worst. Its pixel accuracy is lower than the benchmark pixel accuracy in most cases (when the number of labeled data is 2, 6, 14, or 20). For the density-weighted strategy, its pixel accuracy exceeds the benchmark random sampling sometimes (when the number of labeled data is 2, 6, or 18). However, at other stages (when the number of labeled data is 10, 14, or 20), the pixel accuracy of the density-weighted strategy is lower than the benchmark. For the margin sampling strategy, its pixel accuracy exceeds the benchmark in most cases (when the number of labeled data is 2, 6, 10, or 20). The least confidence strategy performs best compared with other strategies. Its pixel accuracy exceeds the benchmark most times (when the number of labeled data is 2, 6, 10, 14, or 20).

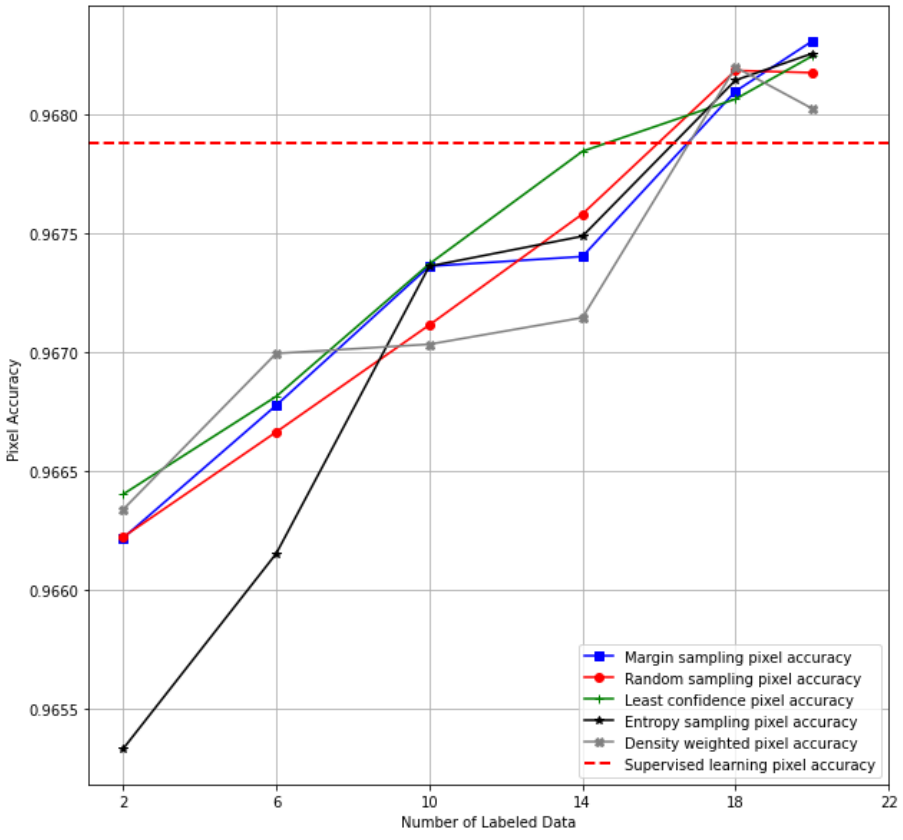
For the CHASEDB1 dataset, one fundus image has two corresponding labels made by two experts, so the CHASEDB1 has two groups of pixel accuracy contrast results. In the subfigure (c) of Figure 19, for the strategies entropy sampling and density-weighted, their changing trends are almost the same, and their pixel accuracy is lower than the benchmark (random sampling) only when the number of labeled data is 8. For strategies least confidence and margin sampling, their pixel accuracy consistently exceeds the benchmark random sampling, the former pixel accuracy also consistently exceeds the latter. Therefore, only the least confidence and margin sampling two strategies perform stably in the current plot.

In the subfigure (d) of Figure 19, when the number of labeled data is 2 or 5, the pixel accuracy of strategies least confidence and entropy sampling exceeds the benchmark random sampling. After that, the pixel accuracy of these two strategies is lower than the benchmark. Then only the strategy least confidence comes back to the level above the benchmark. The strategies density-weighted and margin sampling, whose pixel accuracy consistently exceeds the benchmark in all annotation situations, and the former's pixel accuracy is lower than the latter in most cases (when the number of labeled data is 5, 8, or 11). According to the pixel accuracy comparison between various strategies, only the least confidence and margin sampling two strategies perform stably in the current figure.

Compared with fully supervised learning, the pixel accuracy of active learning strategies has an advantage over the former. For the plots of DRIVE, HRF, and CHASEDB1, when the number of labeled data is equal to or exceeds 14, 18, and 11, the accuracy of active learning strategies surpasses fully supervised learning (see Figure 19).

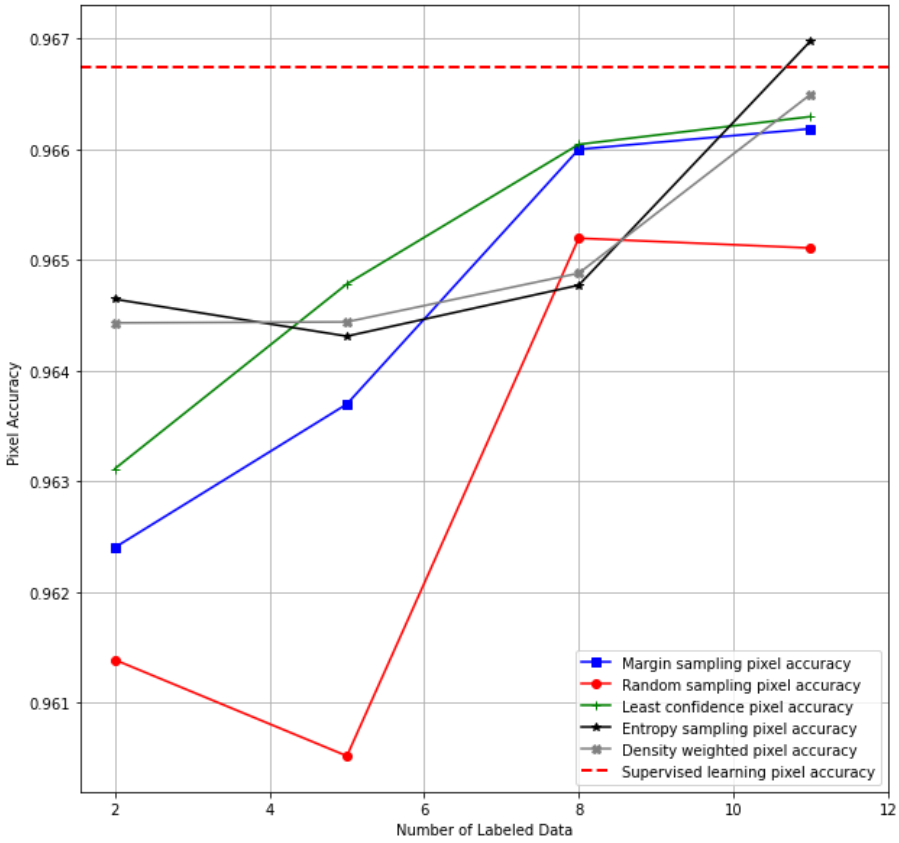


(a)

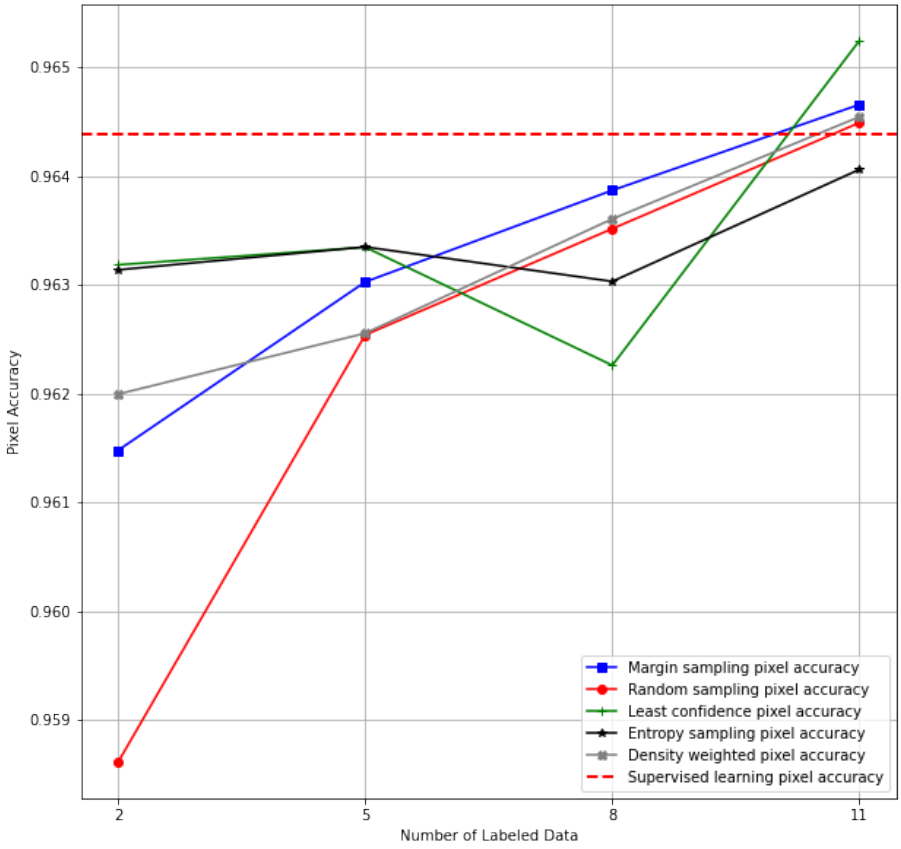


(b)

Figure 19. The contrast of pixel accuracy of various sampling strategies in multiple datasets. (a) DRIVE (b) HRF



(c)



(d)

Figure 19. The contrast of pixel accuracy of various sampling strategies in multiple datasets.(c) CHASEDB1(labels made by the first expert) (d) CHASEDB1(labels made by the second expert).

Meanwhile, related segmentation evaluation indices are also acquired during experiments, which are shown in Table 6, Table 7, Table 8 and Table 9. By reviewing the results in these tables, it is hard to say which strategy performance is better than others on measurement indices sensitivity and specificity. However, with the increase of the annotated images, sensitivity and specificity values are mainly increasing. Even though sometimes their values gradually come down after reaching the highest. For various fundus datasets, comparing sensitivity, F1-score in their tables with their corresponding plots in Figure 19, the changing trend of these two parameters of various strategies is the same as the pixel accuracy, which also explains the variation of some strategies in Figure 19 comes from the sensitivity. Moreover, checking from the general trend of specificity and sensitivity, the latter values are lower than the former. The sensitivity and the specificity commonly refer to TPR and TNR, which represents the actual positive rate and actual negative rate. For the blood vessel segmentation task, positive samples are blood vessel pixels (foreground), negative samples are pixel areas that are not blood vessels (background).

As for the F1-score, its overall trend is also gradually increasing with the number of labeled samples. It shows that the segmentation model's performance is related to the amount of labeled data.

Table 6. Sensitivity, specificity and F1-score of models generated by different annotation situations (DRIVE dataset).

Number of annotated images	Sampling strategies / Evaluation index	Random Sampling	Margin Sampling	Least Confident	Entropy Sampling	Density Weighted
2 images	SEN	**0.6739**	**0.6981**	**0.6551**	**0.7033**	**0.7150**
	SPE	**0.9896**	**0.9874**	**0.9909**	**0.9876**	**0.9855**
	F1	**0.7561**	**0.7634**	**0.7487**	**0.7679**	**0.7664**
6 images	SEN	**0.7282**	**0.7317**	**0.7218**	**0.7213**	**0.7344**
	SPE	**0.9857**	**0.9859**	**0.9869**	**0.9866**	**0.9856**
	F1	**0.7759**	**0.7789**	**0.7770**	**0.7754**	**0.7794**
10 images	SEN	**0.7529**	**0.7510**	**0.7297**	**0.7280**	**0.7196**
	SPE	**0.9836**	**0.9839**	**0.9867**	**0.9864**	**0.9870**
	F1	**0.7827**	**0.7827**	**0.7814**	**0.7790**	**0.7761**
14 images	SEN	**0.7226**	**0.7507**	**0.7330**	**0.7422**	**0.7541**
	SPE	**0.9878**	**0.9853**	**0.9872**	**0.9863**	**0.9852**
	F1	**0.7815**	**0.7887**	**0.7857**	**0.7876**	**0.7904**
18 images	SEN	**0.7226**	**0.7520**	**0.7369**	**0.7301**	**0.7319**
	SPE	**0.9885**	**0.9860**	**0.9875**	**0.9878**	**0.9880**
	F1	**0.7847**	**0.7924**	**0.7894**	**0.7865**	**0.7883**

Table 7. Sensitivity, specifictiy and F1-score of models generated by different annotation situations (CHASEDB1 dataset),whose labels are made by the first expert.

Number of annotated images	Sampling strategies / Evaluation index	Random Sampling	Margin Sampling	Least Confident	Entropy Sampling	Density Weighted
2 images	SEN	**0.7183**	**0.7135**	**0.6748**	**0.6954**	**0.6710**
	SPE	**0.9794**	**0.9808**	**0.9845**	**0.9873**	**0.9862**
	F1	**0.7201**	**0.7241**	**0.7167**	**0.7206**	**0.7229**
5 images	SEN	**0.7389**	**0.7111**	**0.6920**	**0.7012**	**0.7002**
	SPE	**0.9769**	**0.9824**	**0.9850**	**0.9838**	**0.9840**
	F1	**0.7213**	**0.7304**	**0.7310**	**0.7310**	**0.7314**
8 images	SEN	**0.6852**	**0.6741**	**0.6998**	**0.7025**	**0.7167**
	SPE	**0.9859**	**0.9876**	**0.9858**	**0.9842**	**0.9833**
	F1	**0.7313**	**0.7327**	**0.7402**	**0.7339**	**0.7384**
11 images	SEN	**0.7293**	**0.7015**	**0.6998**	**0.7007**	**0.7178**
	SPE	**0.9826**	**0.9858**	**0.9860**	**0.9867**	**0.9849**
	F1	**0.7429**	**0.7415**	**0.7417**	**0.7458**	**0.7476**

Table 8. Sensitivity, specifictiy and F1-score of models generated by different annotation situations (CHASEDB1 dataset),whose labels are made by the second expert.

Number of annotated images	Sampling strategies / Evaluation index	Random Sampling	Margin Sampling	Least Confident	Entropy Sampling	Density Weighted
2 images	SEN	**0.7097**	**0.6769**	**0.6577**	**0.6186**	**0.6055**
	SPE	**0.9771**	**0.9826**	**0.9858**	**0.9887**	**0.9884**
	F1	**0.7034**	**0.7085**	**0.7119**	**0.6989**	**0.6878**
5 images	SEN	**0.6913**	**0.6747**	**0.6524**	**0.6646**	**0.6608**
	SPE	**0.9826**	**0.9844**	**0.9864**	**0.9855**	**0.9849**
	F1	**0.7185**	**0.7162**	**0.7111**	**0.7149**	**0.7094**
8 images	SEN	**0.7064**	**0.6727**	**0.6980**	**0.7144**	**0.6851**
	SPE	**0.9826**	**0.9854**	**0.9818**	**0.9814**	**0.9842**
	F1	**0.7281**	**0.7203**	**0.7189**	**0.7277**	**0.7225**
11 images	SEN	**0.7206**	**0.6789**	**0.6929**	**0.7198**	**0.7078**
	SPE	**0.9826**	**0.9858**	**0.9854**	**0.9822**	**0.9836**
	F1	**0.7373**	**0.7265**	**0.7338**	**0.7347**	**0.7341**

Table 9. Sensitivity, specifictiy and F1-score of models generated by different annotation situations (HRF dataset).

Number of annotated images	Sampling strategies / Evaluation index	Random Sampling	Margin Sampling	Least Confident	Entropy Sampling	Density Weighted
2 images	SEN	0.7544	0.7313	0.7364	0.7701	0.7302
	SPE	0.9842	0.9862	0.9859	0.9819	0.9864
	F1	0.7779	0.7725	0.7746	0.7770	0.7728
6 images	SEN	0.7348	0.7018	0.7165	0.7014	0.7290
	SPE	0.9863	0.9893	0.9881	0.9886	0.9872
	F1	0.7756	0.7681	0.7720	0.7647	0.7760
10 images	SEN	0.7419	0.7392	0.7488	0.7133	0.7544
	SPE	0.9862	0.9867	0.9859	0.9889	0.9851
	F1	0.7796	0.7803	0.7825	0.7741	0.7821
14 images	SEN	0.7261	0.7172	0.7162	0.7368	0.6948
	SPE	0.9881	0.9886	0.9892	0.9871	0.9903
	F1	0.7784	0.7753	0.7774	0.7804	0.7683
18 images	SEN	0.7395	0.7305	0.7641	0.7424	0.7541
	SPE	0.9876	0.9883	0.9854	0.9873	0.9864
	F1	0.7847	0.7821	0.7896	0.7852	0.7880
20 images	SEN	0.7211	0.7378	0.7305	0.7443	0.7443
	SPE	0.9891	0.9879	0.9884	0.9873	0.9870
	F1	0.7804	0.7850	0.7829	0.7862	0.7849

Next, the experiment uses testing datasets of three fundus datasets as the validation dataset, aiming to evaluate the model change in pixel accuracy during the training process. The current study discusses the most typical case of less labeled data learning: only two fundus images are labeled for each dataset. One thing needs to be explained here, the number of patches is calculated by:

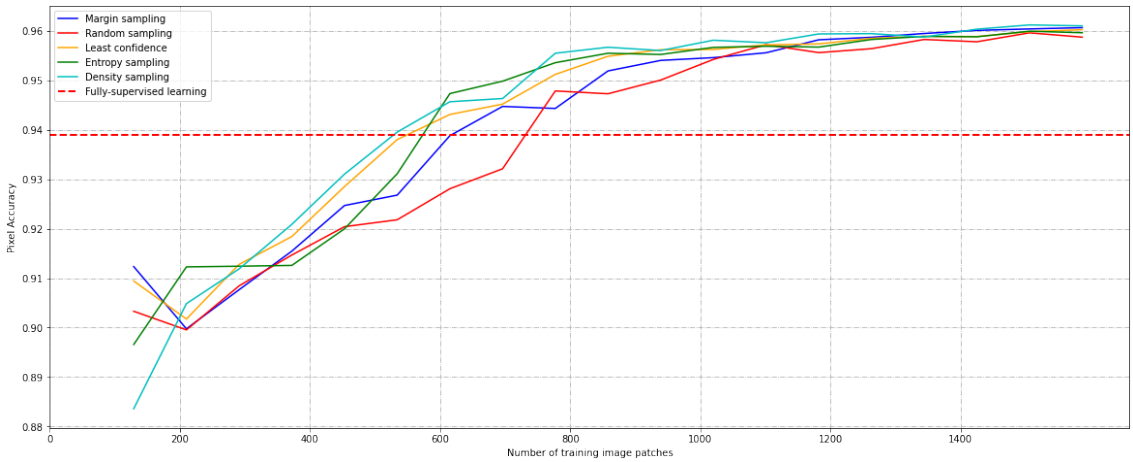
$$N_{\text{patch}} = N_{\text{patch\_per\_image}} N_{\text{image}} \tag{17}$$

where $N_{\text{patch\_per\_image}}$ is the number of patches generated for each fundus image, which comes from using specified patch size to average the resized image. N_{image} is the number of images from divided fundus datasets (see the Table 5). Related parameters setting information about the current experiment are displayed in Table 10.

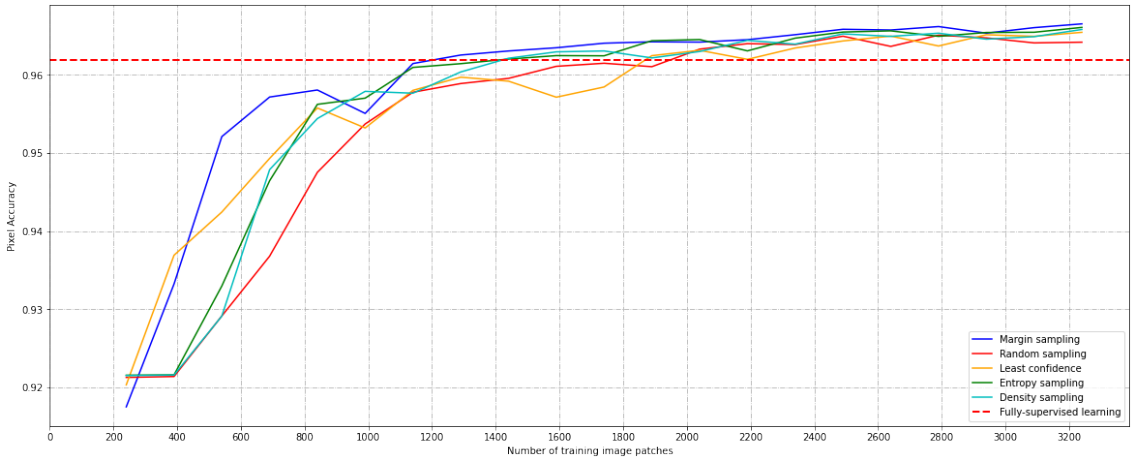
Table 10. Related settings of experimental parameters when annotated number = 2 (DRIVE dataset). N_{active}-number of patches generated from annotated images in active learning. N_{valid}-number of patches generated from the validation dataset. $N_{\text{supervised}}$-number of patches generated from annotated samples in fully-supervised learning.

Database	Resized resolution	Patch size	N_{active}	N_{valid}	$N_{\text{supervised}}$
DRIVE	576x576	64x64	65	1620	1620
HRF	1440x960	96x96	240	3450	3300
CHASEDB1	960x960	96x96	160	1400	1400

Figure 20 describes the trend of testing accuracy with the increase of training image patches for three fundus datasets. For the subfigures (a) and (b) of Figure 20, it can be seen that the fully supervised learning pixel accuracy of DRIVE is lower than HRF, that is because the training patches from the latter have better quality than the former. These two figures also show one result: active learning strategies can use fewer image patches and less time to get better accuracy in the HRF dataset. For the subfigures (c) and (d) of Figure 20, their plotting results are from the training model using annotations of two experts. Plotting results show that the pixel accuracy of all strategies is upward with the increasing number of patches. However, it can only get close to the accuracy of fully supervised learning. Comparing these two plots and the HRF plot, active learning strategies spend more time and training patches to reach higher accuracy.

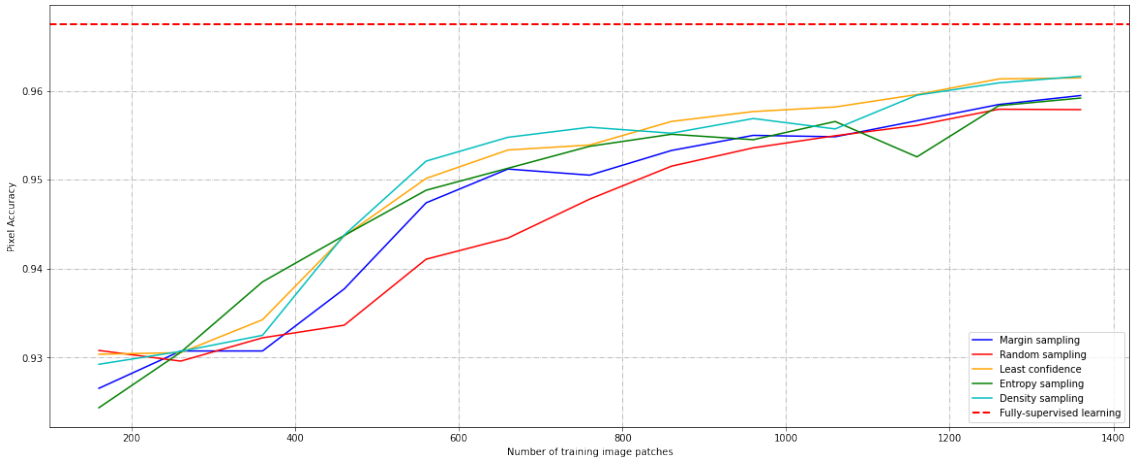


(a)

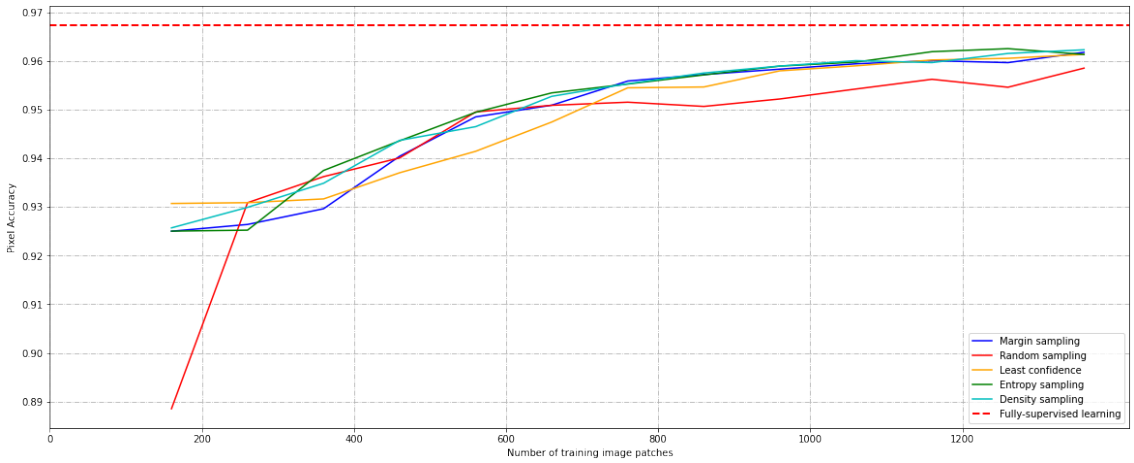


(b)

Figure 20. Pixel accuracy as the function of the number of training image patches: (a) DRIVE (b) HRF



(c)

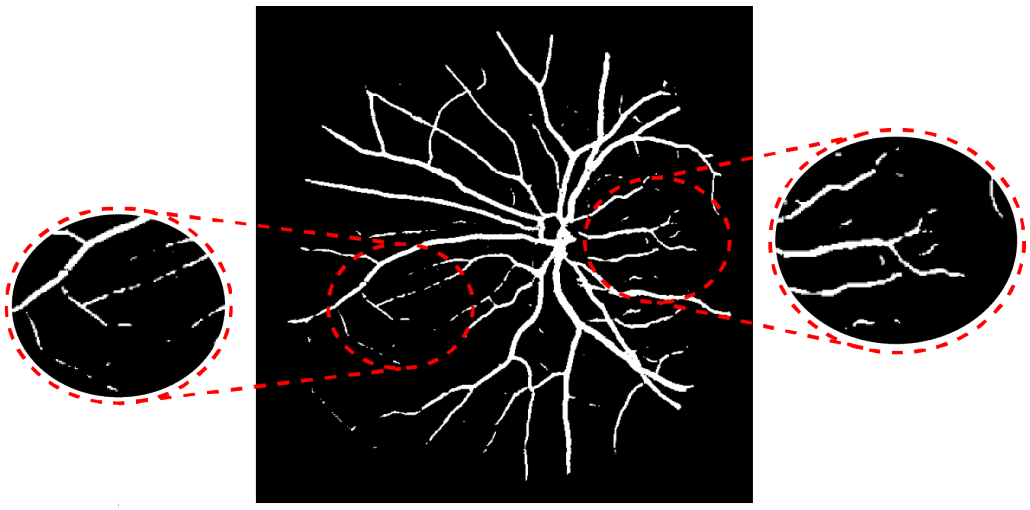


(d)

Figure 20. Pixel accuracy as the function of the number of training image patches: (c) CHASEDB1 (Annotations made by the first expert). (d) CHASEDB1 (Annotations made by the second expert).

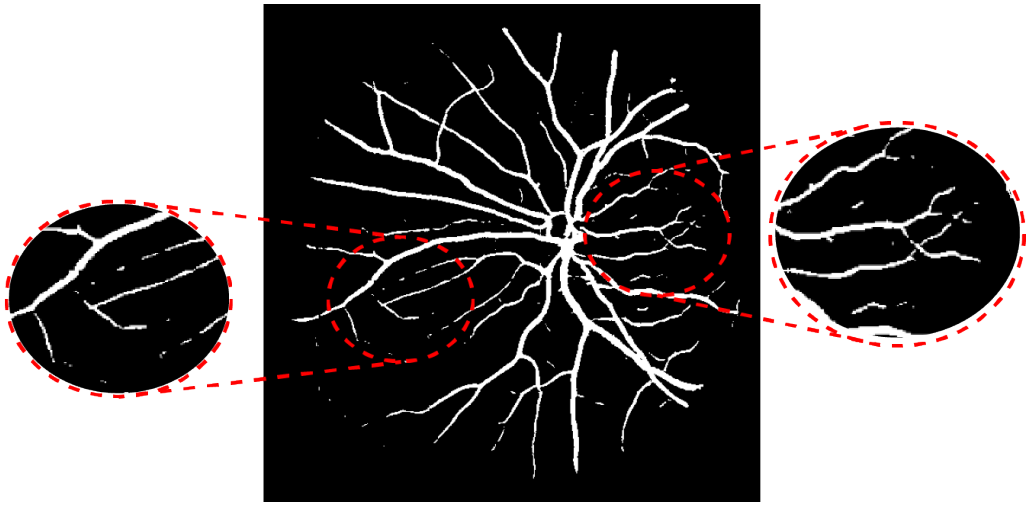
According to the pixel accuracy performance of various strategies in Figure 19, for each fundus dataset, the current study selects the segmentation results of a strategy with a stable pixel accuracy change for display. As seen in Figures 21, 22, 23, and 24, when the number of annotated images reaches a certain level, the segmentation results of active learning strategies are better than or get close to fully supervised learning. For these four figures, the level values are 14, 18, 8, and 8 (corresponding to figures 21 (f), 22 (f), 23 (d), and 24 (d)), respectively. Besides, for the CHASEDB1 dataset, active learning can get better segmentation results of tiny vessels from the second expert annotations (see Figures 23 and 24).

(a)

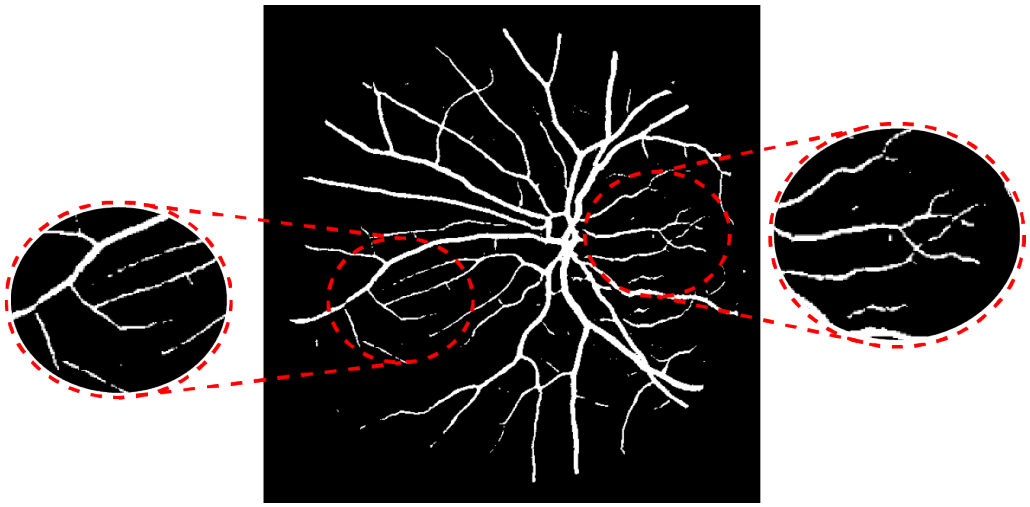


(b)

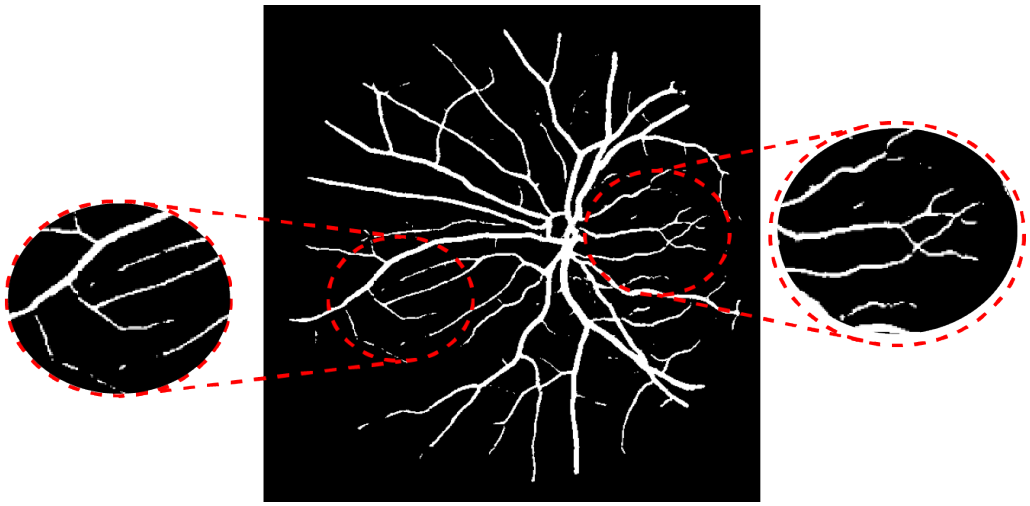
Figure 21. Density-weighted segmentation results(DRIVE dataset): (a) The image to be segmented (b) The segmentation results when annotated number N = 2.



(c)

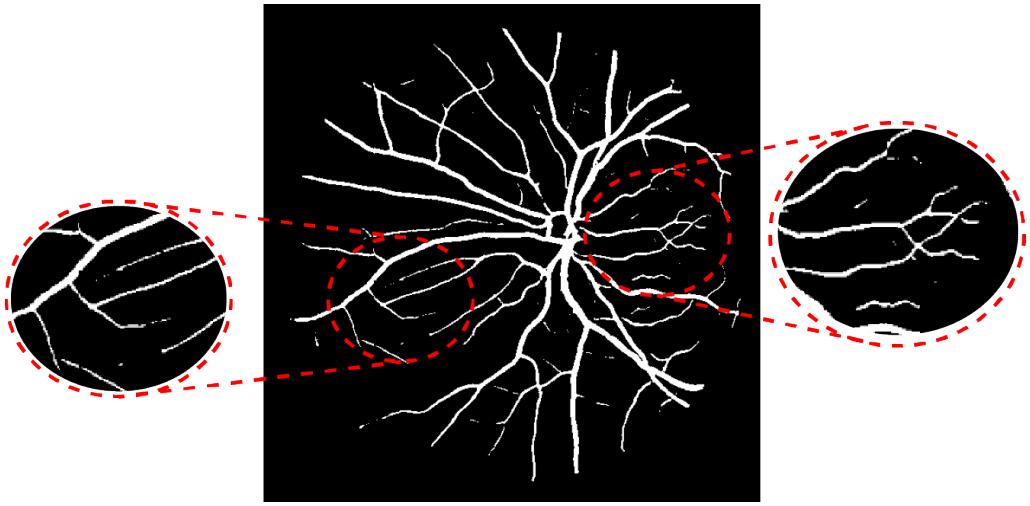


(d)

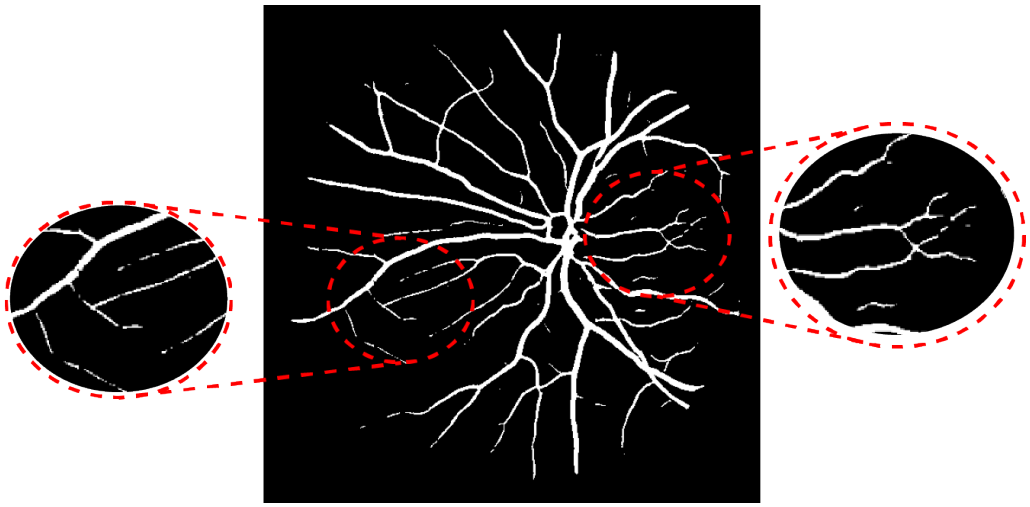


(e)

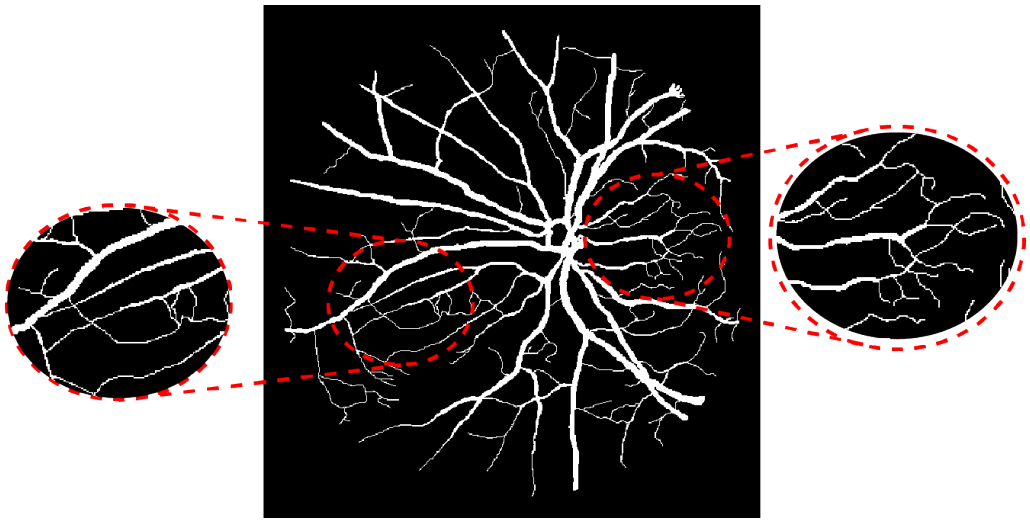
Figure 21. Density-weighted segmentation results(DRIVE dataset): (c) The segmentation results when annotated number N = 6 (d) The segmentation results when annotated number N = 10 (e) The segmentation results when annotated number N = 14.



(f)

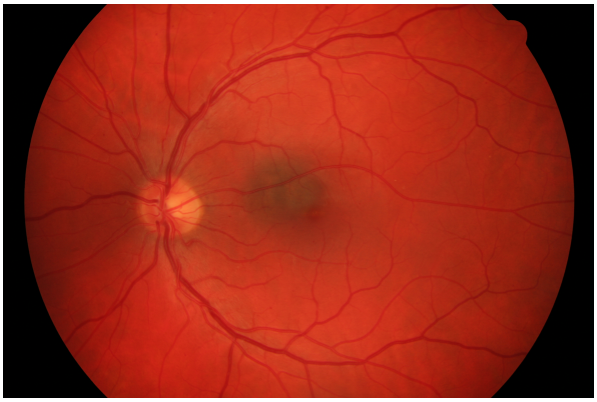


(g)

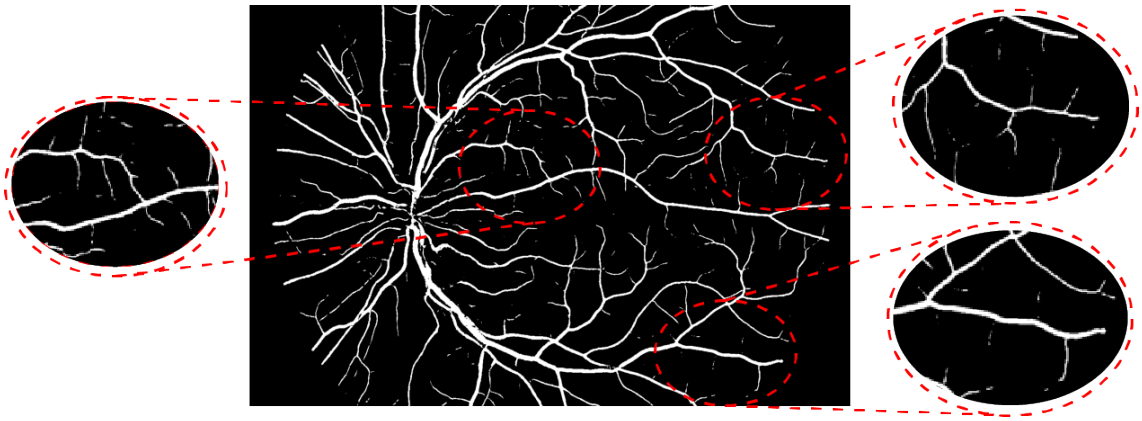


(h)

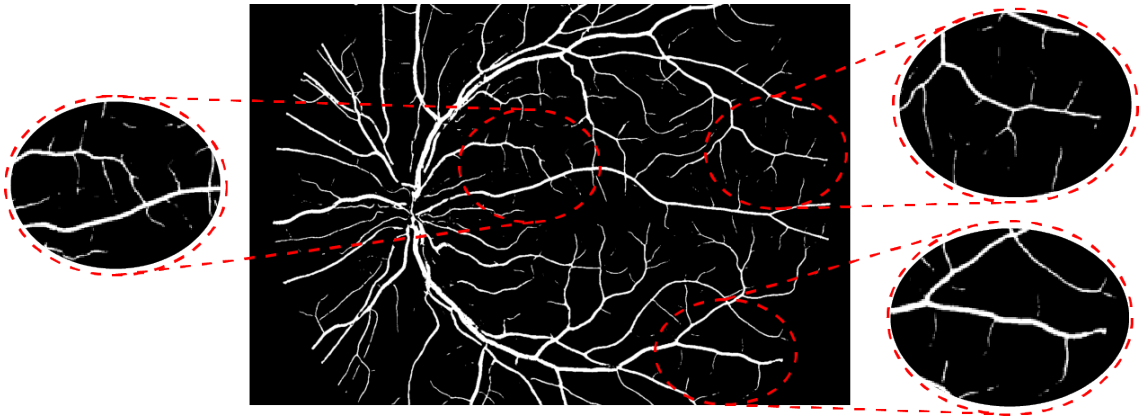
Figure 21. Density-weighted segmentation results(DRIVE dataset): (f) The segmentation results when annotated number N = 18. (g) The segmentation results from fully supervised learning (h) The groundtruth of the current fundus image.



(a)

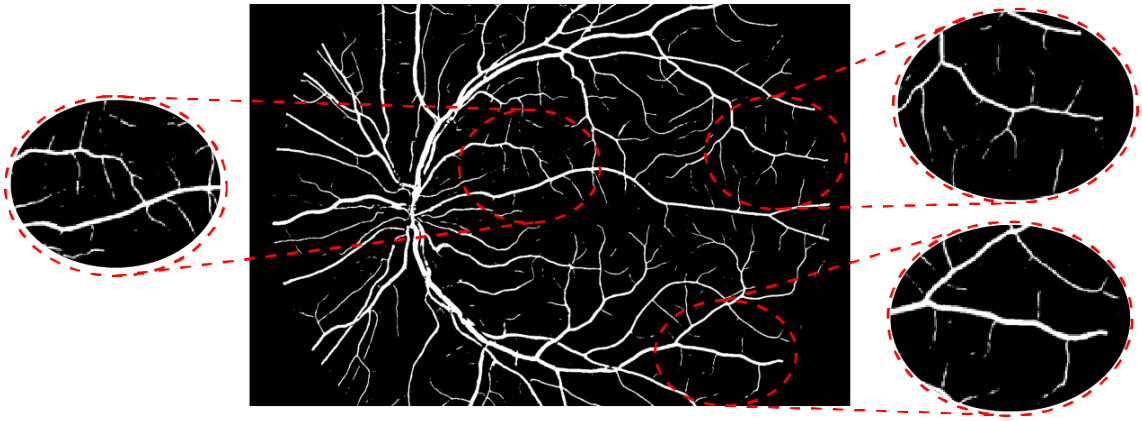


(b)

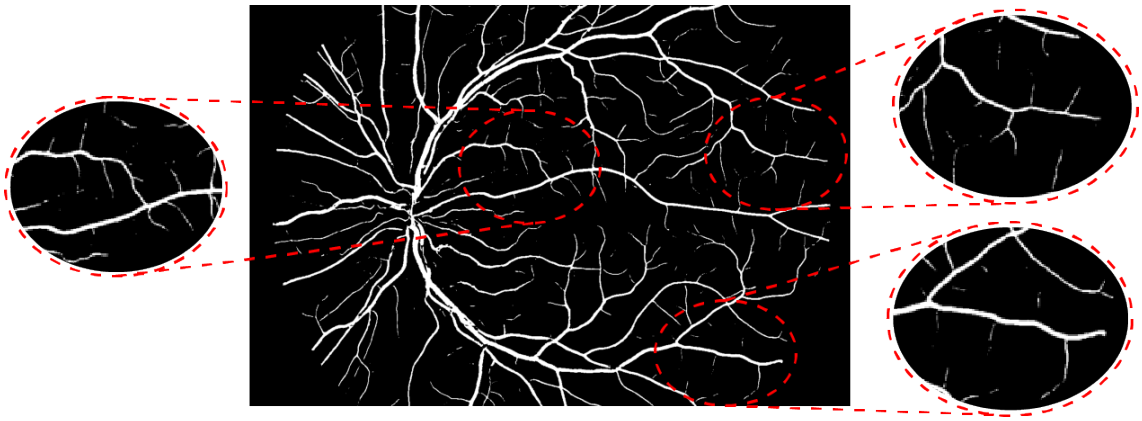


(c)

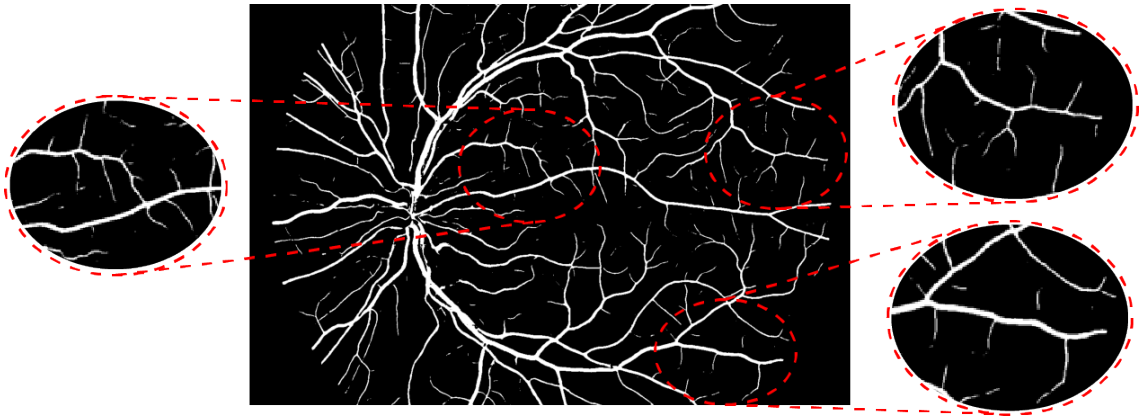
Figure 22. Least confidence segmentation results(HRF dataset): (a) The image to be segmented (b) The segmentation results when annotated number N = 2. (c) The segmentation results when annotated number N = 6.



(d)

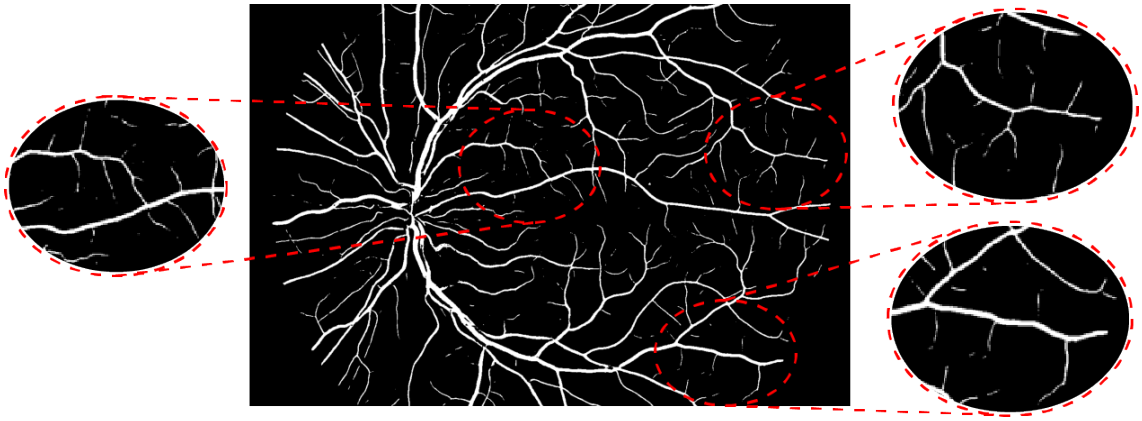


(e)

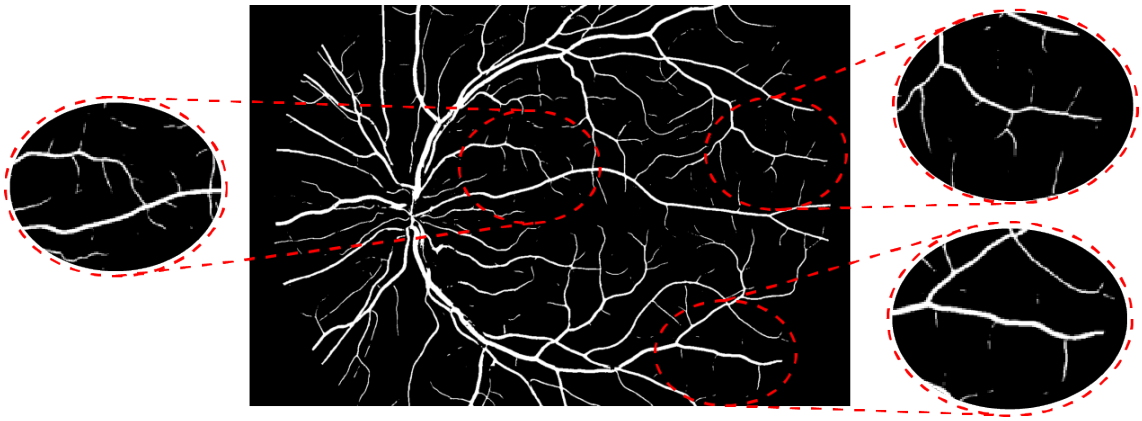


(f)

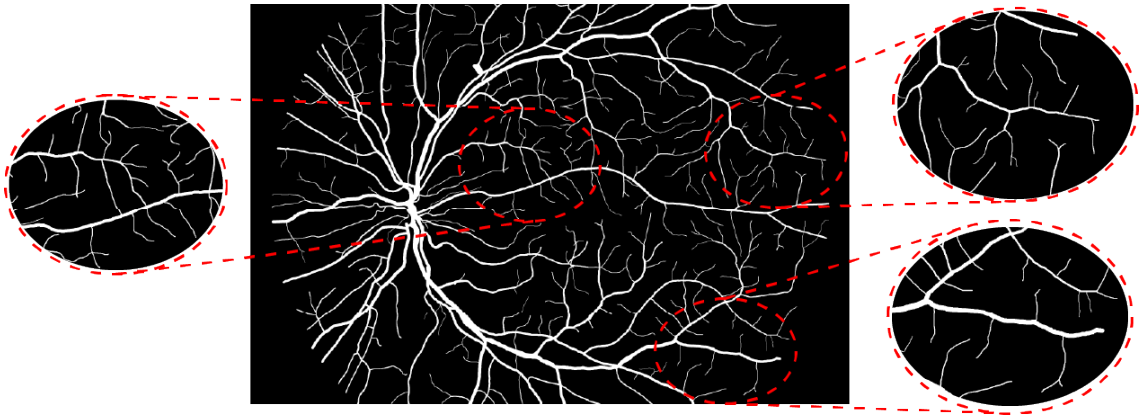
Figure 22. Least confidence segmentation results(HRF dataset): (d) The segmentation results when annotated number N = 10 (e) The segmentation results when annotated number N = 14 (f) The segmentation results when annotated number N = 18.



(g)

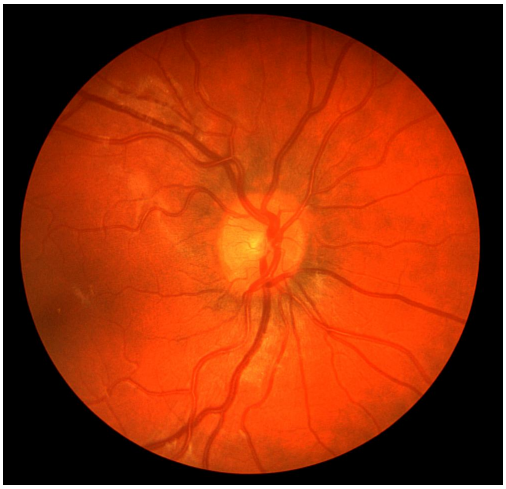


(h)

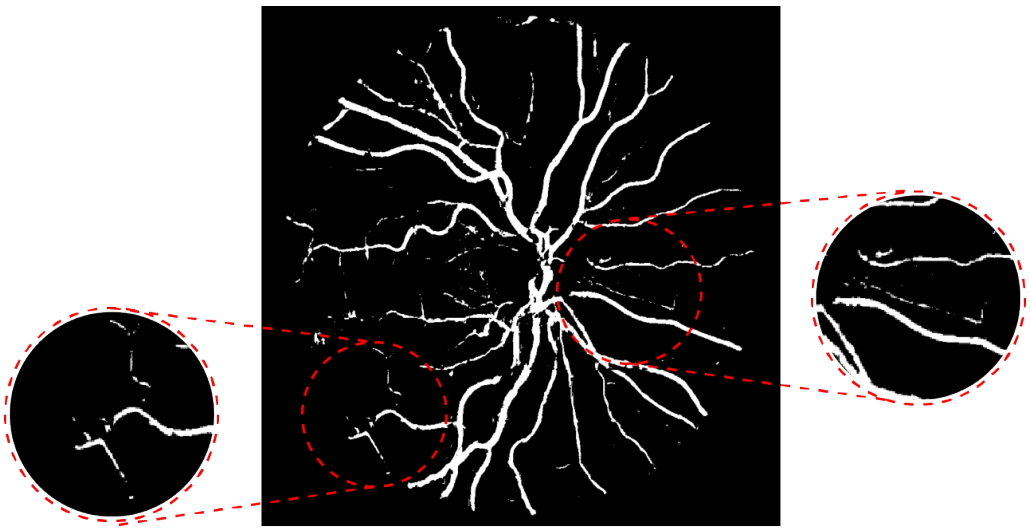


(i)

Figure 22. Least confidence segmentation results(HRF dataset): (g) The segmentation results when annotated number N = 20. (h) The segmentation results from fully supervised learning (i) The groundtruth of the current fundus image.

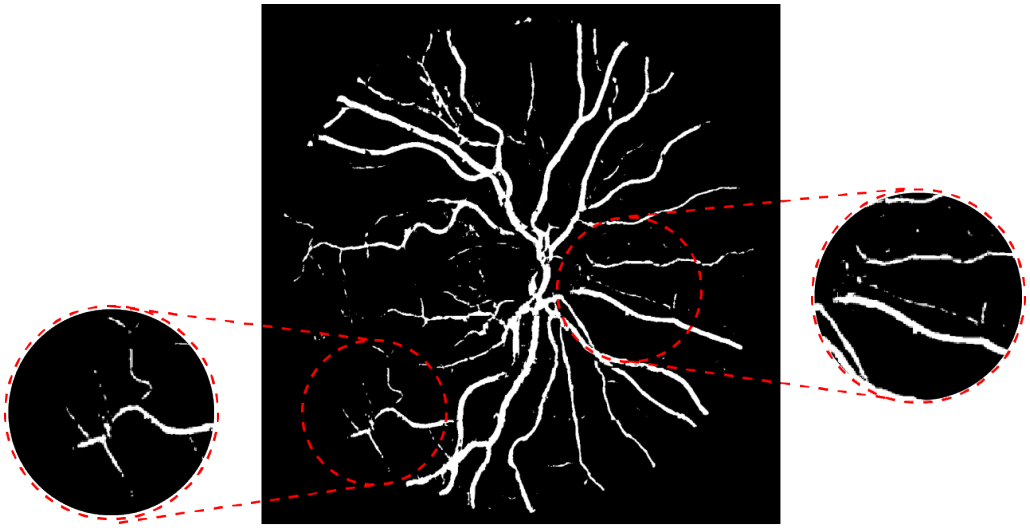


(a)

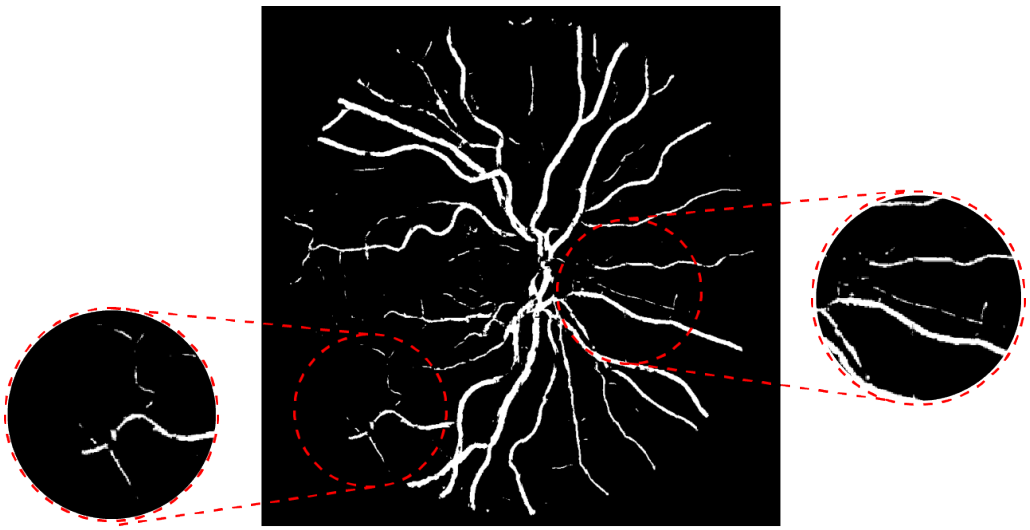


(b)

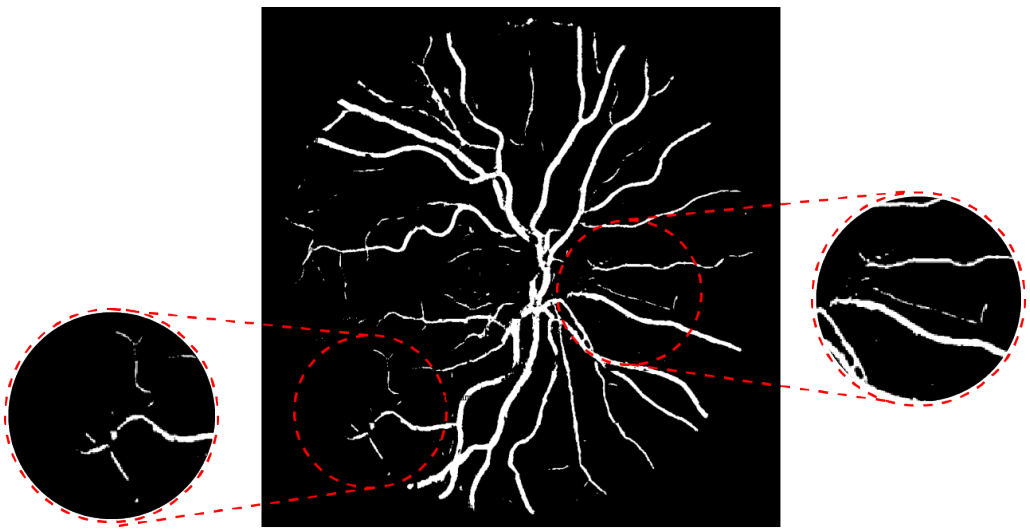
Figure 23. Margin sampling segmentation results (CHASEDB1 dataset, and its annotations come from the first expert): (a) The image to be segmented (b) The segmentation results when annotated number N = 2.



(c)

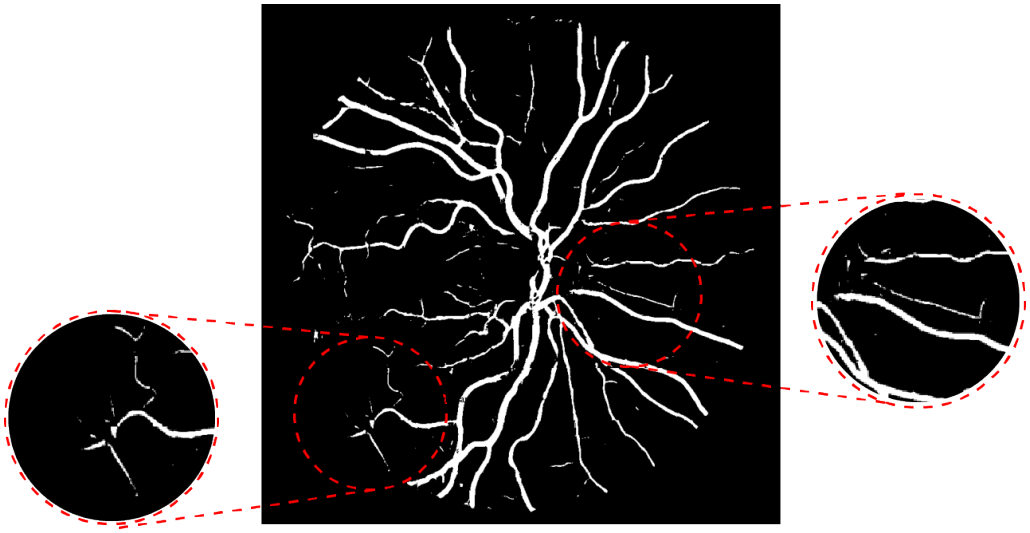


(d)

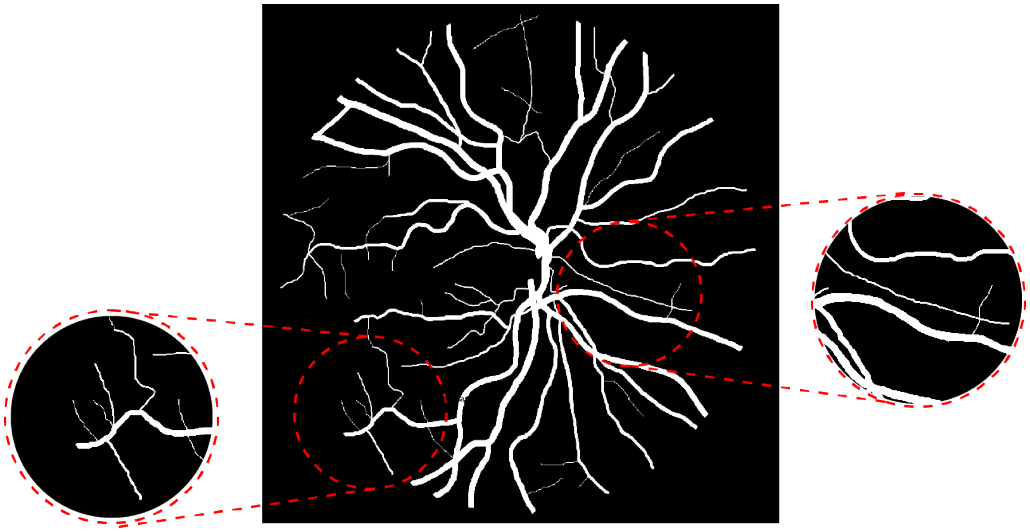


(e)

Figure 23. Margin sampling segmentation results (CHASEDB1 dataset, and its annotations come from the first expert): (c) The segmentation results when annotated number N = 5. (d) The segmentation results when annotated number N = 8 (e) The segmentation results when annotated number N = 11.

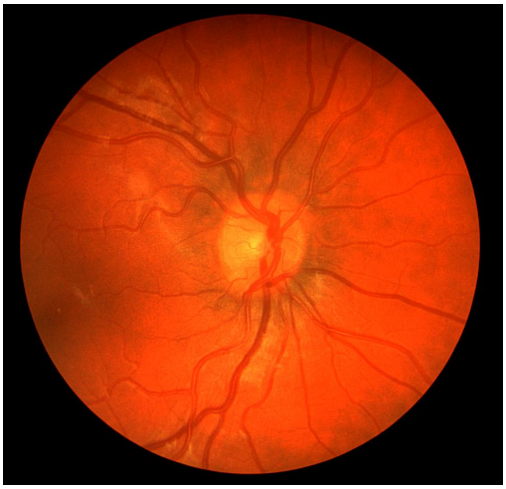


(f)

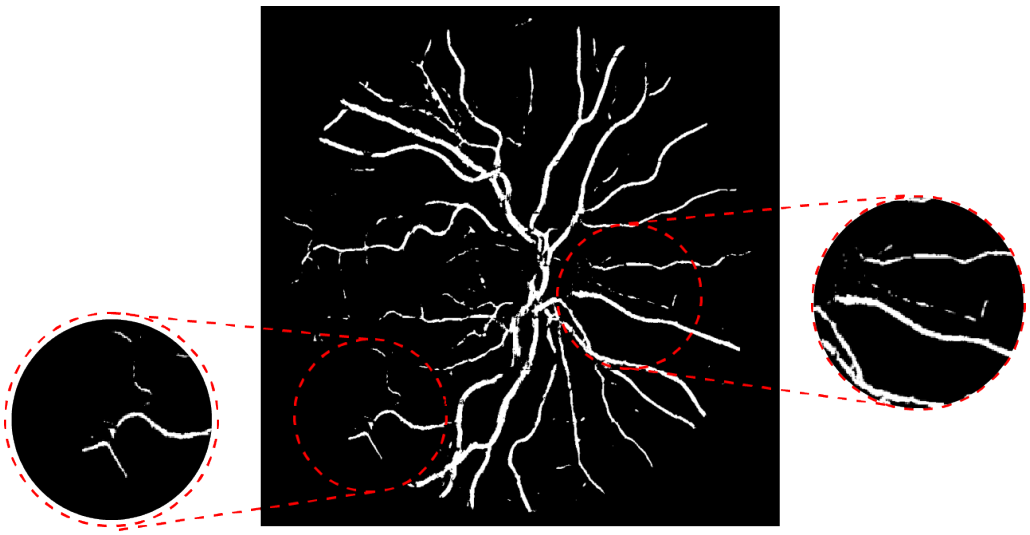


(g)

Figure 23. Margin sampling segmentation results (CHASEDB1 dataset, and its annotations come from the first expert): (f) The segmentation results from fully supervised learning (g) The groundtruth of the current fundus image.

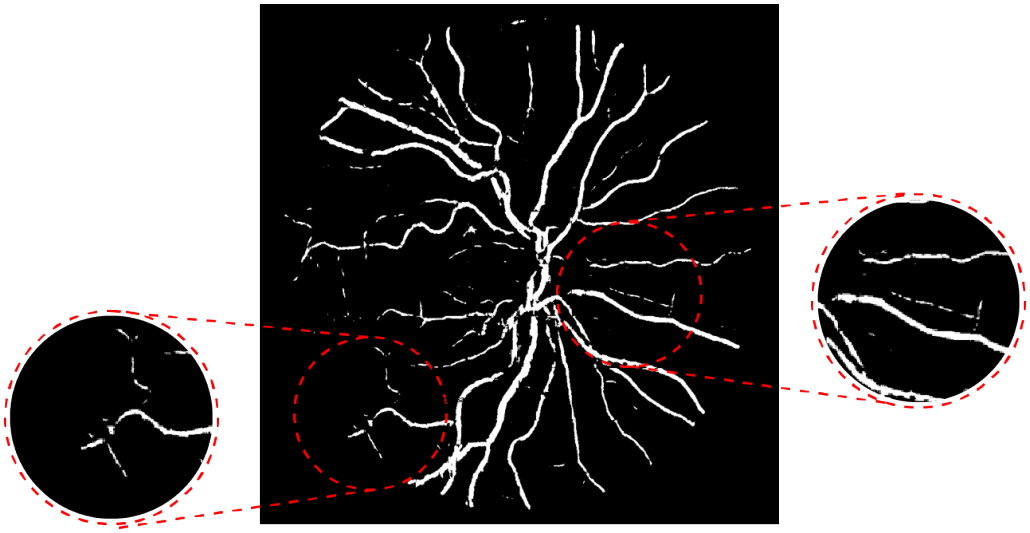


(a)

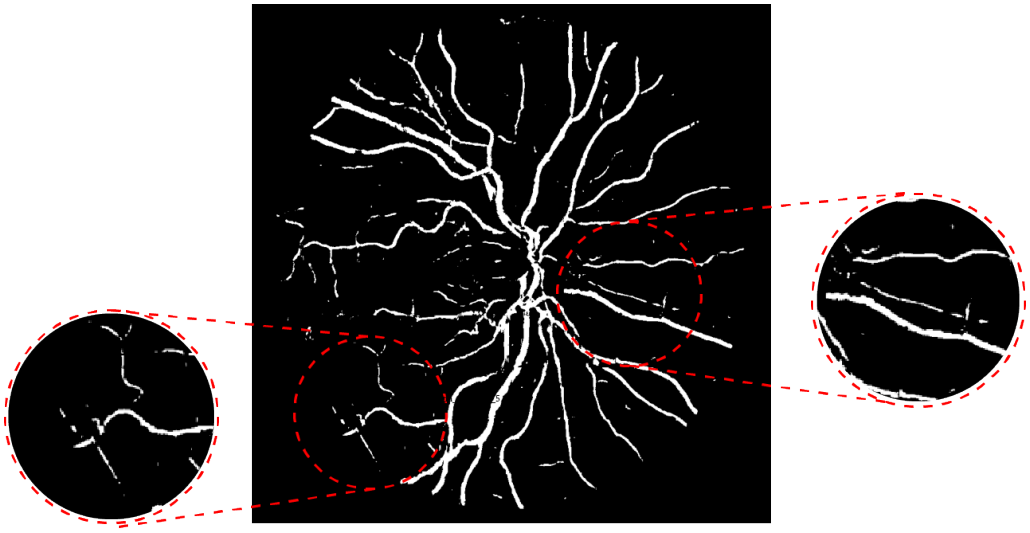


(b)

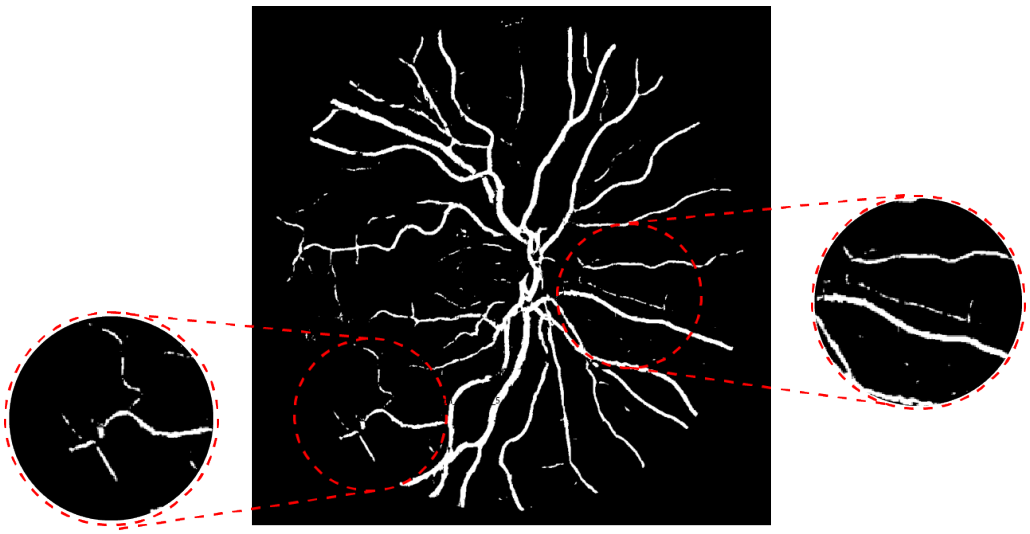
Figure 24. Margin sampling segmentation results (CHASEDB1 dataset, and its annotations come from the second expert): (a) The image to be segmented (b) The segmentation results when annotated number N = 2.



(c)

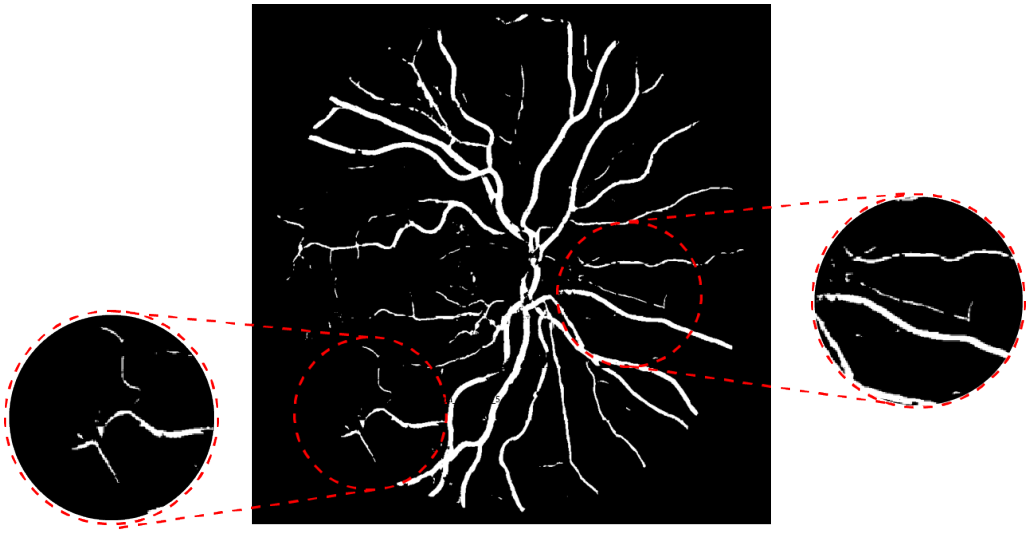


(d)

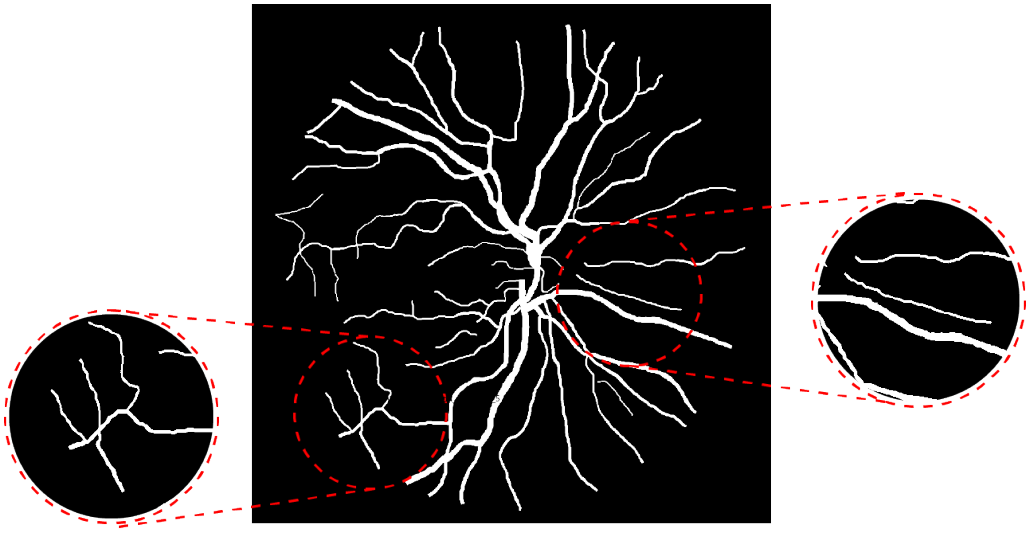


(e)

Figure 24. Margin sampling segmentation results (CHASEDB1 dataset, and its annotations come from the second expert): (c) The segmentation results when annotated number N = 5 (d) The segmentation results when annotated number N = 8 (e) The segmentation results when annotated number N = 11.



(f)



(g)

Figure 24. Margin sampling segmentation results (CHASEDB1 dataset, and its annotations come from the second expert): (f) The segmentation results from fully supervised learning (g) The groundtruth of the current fundus image.

Making a comparison between the GT and segmentation results in various datasets, the active learning performs better for the segmentation of narrow vessels in DRIVE (see the subfigure (e) of Figure 21). For the segmentation of narrow vessels in the HRF dataset, active learning needs to use more (annotated number N = 18) labeled samples than DRIVE to get the best results (see the subfigure (f) of Figure 22). For the segmentation results of narrow vessels in the CHASEDB1 dataset, active learning performs worse than the previous two datasets (see the subfigures (d) of Figure 23).

5 DISCUSSION

In the absence of annotated data, the current research studies contributions of various active learning strategies to the model performance, thereby obtaining the corresponding conclusions. However, some aspects of the algorithm are still worth discussing. The following content will discuss from two angles: the research significance of the algorithm and places where the algorithm needs to be improved.

5.1 Results of the current study

According to the experimental results shown in Chapter 4, most proposed active learning strategies behave well according to pixelwise accuracy. With the increasing number of training image patches and annotated images, most strategy accuracies exceed the baseline (random sampling) during the training and testing process in most cases (see Figure 19). When the amount of annotated data reaches a specific level, active learning's segmentation pixelwise accuracy is close or even higher than fully supervised learning (see Figure 19). From these perspectives, active learning can improve segmentation performance indeed.

For the current implemented algorithm, traditional image preprocessing helped to improve the fundus image's quality, especially the CLAHE algorithm, which makes blood vessels more clear to identify after processing. Then it comes to the segmentation network. In order to save time cost used for training, the author does not use the whole U-net as the model architecture. To make the training of the blood vessel segmentation method more efficient, the whole U-Net architecture was not used. Instead, two downsampling layers and two upsampling layers are removed from the original network, adding enough BatchNormalization layers to ensure that segmentation accuracy is not decreased, which has caused the network not to learn deeper semantic information from images. Therefore, some tiny vessels in the GT do not show up in segmentation results. Another improvement of this algorithm is that it adopts the combination of pre-training and continuously fine-tuning. This combination can improve the model performance significantly when getting fewer data annotated, but it also causes that active learning benefits in the current study are not very obvious.

Many challenges also happen in the process of implementation. For active learning strategies, confidence generally needs the calculation to identify the next candidate. However,

the confidence calculation is for each pixel of one image. In the thesis project, the entire image confidence is acquired by taking the average of every pixel confidence. The author analyzes that the average cannot represent the whole image confidence precisely. Besides, the imbalance between the positive and negative samples also brings some issues. The proportion of negative samples(background) is higher than the positive samples(blood vessels), which causes the high training accuracy to show up from the start of training the pre-trained model.

5.2 Future work

Given the problems raised above, the algorithm can be improved in the following aspects: to begin with, for identifying more precise confidence of an image. Other active learning strategies like QBC, Excepted Model Change (EMC), and Excepted Error Reduction (EER) can be taken into consideration. The next part that needs improvement is the imbalance between the classes. One opinion is to utilize other loss functions (like Dice [84]) to suppress the negative samples.

Moreover, a more challenging task such as A/V classification could be used for further study. When training a pre-trained pre-trained model of blood vessel segmentation, its final training accuracy generally is very high. It is because the model must get plenty of semantic information from the start. However, it might be different for the A/V classification task. Features of arteries and veins generally are challenging to learn by the model. Therefore, one multi-task network for solving both blood vessel segmentation and A/V classification is worthy of implementation in the future, and there are at least two possible solutions that could be used to make it a reality. The first one is proposed by Xu et al. in [85], to realize A/V classification, they propose a loss function used for training, whose cost is decided by assigning arteries and veins weights according to their proportion in the image. Another one is proposed by Li et al. in [86]. They develop a two-step vessel classification named as SeqNet, which is based on U-net. Its first step is to get the blood vessel segmentation results, and the second step is to implement A/V classification based on segmentation results from the first step.

6 CONCLUSION

This thesis mainly studies the role of active learning strategies in training retinal image analysis methods. With limited fundus images labeled for the blood vessel segmentation task, active learning strategies including margin sampling, least confidence, entropy sampling, and density-weighted method are implemented to carry out unlabeled sample selection. The experimental results show that the segmentation accuracy of the model trained through the active learning strategy exceeds the benchmark, that is, random sampling. When the annotated number of images reaches a certain level, the active learning segmentation pixelwise accuracy gets close to or exceeds the fully supervised learning. The former requires fewer samples for training in this situation.

REFERENCES

[1] W Todd Cade. Diabetes-related microvascular and macrovascular diseases in the physical therapy setting. *Physical therapy*, 88(11):1322–1335, 2008.

[2] Oliver Faust, Rajendra Acharya, Eddie Yin-Kwee Ng, Kwan-Hoong Ng, and Jasjit S Suri. Algorithms for the automated detection of diabetic retinopathy using digital fundus images: a review. *Journal of medical systems*, 36(1):145–157, 2012.

[3] Wang C Pang H. Deep learning model for diabetic retinopathy detection. *Software S O.*, 028(011):3018–3029, 2017.

[4] Yehui Yang, Tao Li, Wensi Li, Haishan Wu, Wei Fan, and Wensheng Zhang. Lesion detection and grading of diabetic retinopathy via two-stages deep convolutional neural networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 533–540. Springer, 2017.

[5] Clara I Sánchez, Meindert Niemeijer, Michael D Abràmoff, and Bram van Ginneken. Active learning for an efficient training strategy of computer-aided diagnosis systems: application to diabetic retinopathy screening. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 603–610. Springer, 2010.

[6] R. Monarch. *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI*. Manning Publications, 2021.

[7] Burr Settles. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2009.

[8] Zongwei Zhou, Jae Shin, Lei Zhang, Suryakanth Gurudu, Michael Gotway, and Jianming Liang. Fine-tuning convolutional neural networks for biomedical image analysis: actively and incrementally. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7340–7351, 2017.

[9] Jia-Jie Zhu and José Bento. Generative adversarial active learning. *arXiv preprint arXiv:1702.07956*, 2017.

[10] Ksenia Konyushkova, Raphael Sznitman, and Pascal Fua. Learning active learning from data. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[11] Tomi Kauppi. *Eye fundus image analysis for automatic detection of diabetic retinopathy*. PhD thesis, Lappeenranta University of Technology, 2010.

[12] Aswani Dutt Vadlapudi, Ashaben Patel, Kishore Cholkar, and Ashim K Mitra. Recent patents on emerging therapeutics for the treatment of glaucoma, age related macular degeneration and uveitis. *Recent patents on biomedical engineering*, 5(1):83, 2012.

[13] Helga Kolb. How the retina works: Much of the construction of an image takes place in the retina itself through the use of specialized neural circuits. *American scientist*, 91(1):28–35, 2003.

[14] Norah Asiri, Muhammad Hussain, Fadwa Al Adel, and Nazih Alzaidi. Deep learning based computer-aided diagnosis systems for diabetic retinopathy: A survey. *Artificial intelligence in medicine*, 99:101701, 2019.

[15] Early Treatment Diabetic Retinopathy Study Research Group et al. Early photocoagulation for diabetic retinopathy: Etdrs report number 9. *Ophthalmology*, 98(5):766–785, 1991.

[16] CP Wilkinson, Frederick L Ferris III, Ronald E Klein, Paul P Lee, Carl David Agardh, Matthew Davis, Diana Dills, Anselm Kampik, R Pararajasegaram, Juan T Verdaguer, et al. Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. *Ophthalmology*, 110(9):1677–1682, 2003.

[17] Rishab Gargeya and Theodore Leng. Automated identification of diabetic retinopathy using deep learning. *Ophthalmology*, 124(7):962–969, 2017.

[18] AD Hoover, Valentina Kouznetsova, and Michael Goldbaum. Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. *IEEE Transactions on Medical imaging*, 19(3):203–210, 2000.

[19] Joes Staal, Michael D Abràmoff, Meindert Niemeijer, Max A Viergever, and Bram Van Ginneken. Ridge-based vessel segmentation in color images of the retina. *IEEE transactions on medical imaging*, 23(4):501–509, 2004.

[20] Tomi Kauppi, Valentina Kalesnykiene, Joni-Kristian Kamarainen, Lasse Lensu, Iiris Sorri, Hannu Uusitalo, Heikki Kälviäinen, and Juhani Pietilä. Diaretdb0: Evaluation database and methodology for diabetic retinopathy algorithms. *Machine Vision and Pattern Recognition Research Group, Lappeenranta University of Technology, Finland*, 73:1–17, 2006.

[21] Tomi Kauppi, Valentina Kalesnykiene, Joni-Kristian Kamarainen, Lasse Lensu, Iiris Sorri, Asta Raninen, Raija Voutilainen, Hannu Uusitalo, Heikki Kälviäinen, and Juhani Pietilä. The diaretdb1 diabetic retinopathy database and evaluation protocol. In *The British Machine Vision Conference (BMVC)*, volume 1, pages 1–10, 2007.

[22] Etienne Decencière, Xiwei Zhang, Guy Cazuguel, Bruno Lay, Béatrice Cochener, Caroline Trone, Philippe Gain, Richard Ordonez, Pascale Massin, Ali Erginay, et al. Feedback on a publicly distributed image database: the messidor database. *Image Analysis & Stereology*, 33(3):231–234, 2014.

[23] Bashir Al-Diri, Andrew Hunter, David Steel, Maged Habib, Taghread Hudaib, and Simon Berry. A reference data set for retinal vessel profiles. In *30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 2262–2265. IEEE, 2008.

[24] Meindert Niemeijer, Bram Van Ginneken, Michael J Cree, Atsushi Mizutani, Gwénolé Quellec, Clara I Sánchez, Bob Zhang, Roberto Hornero, Mathieu Lamard, Chisako Muramatsu, et al. Retinopathy online challenge: automatic detection of microaneurysms in digital color fundus photographs. *IEEE transactions on medical imaging*, 29(1):185–195, 2009.

[25] Luca Giancardo, Fabrice Meriaudeau, Thomas P Karnowski, Yaquin Li, Kenneth W Tobin, and Edward Chaum. Automatic retina exudates segmentation without a manually labelled training set. In *IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 1396–1400. IEEE, 2011.

[26] Xiwei Zhang, Guillaume Thibault, Etienne Decencière, Beatriz Marcotegui, Bruno Laÿ, Ronan Danno, Guy Cazuguel, Gwénolé Quellec, Mathieu Lamard, Pascale Massin, et al. Exudate detection in color retinal images for mass screening of diabetic retinopathy. *Medical image analysis*, 18(7):1026–1043, 2014.

[27] Adarsh Pradhan, Bhaskarjyoti Sarma, Rahul Kumar Nath, Ajay Das, and Anirudha Chakraborty. Diabetic retinopathy detection on retinal fundus images using convolutional neural network. In *International Conference on Machine Learning, Image Processing, Network Security and Data Sciences*, pages 254–266. Springer, 2020.

[28] Christopher G Owen, Alicja R Rudnicka, Robert Mullen, Sarah A Barman, Dorothy Monekosso, Peter H Whincup, Jeffrey Ng, and Carl Paterson. Measuring retinal vessel tortuosity in 10-year-old children: validation of the computer-assisted image analysis of the retina (caiar) program. *Investigative ophthalmology & visual science*, 50(5):2004–2010, 2009.

[29] J Odstrčilík, Jiri Jan, J Gazárek, and R Kolář. Improvement of vessel segmentation by matched filtering in colour retinal images. In *World Congress on Medical Physics and Biomedical Engineering, September 7-12, 2009, Munich, Germany*, pages 327–330. Springer, 2009.

[30] Qiaoliang Li, Bowei Feng, LinPei Xie, Ping Liang, Huisheng Zhang, and Tianfu Wang. A cross-modality learning approach for vessel segmentation in retinal images. *IEEE transactions on medical imaging*, 35(1):109–118, 2015.

[31] Ning Du and Yafen Li. Automated identification of diabetic retinopathy stages using support vector machine. In *Proceedings of the 32nd Chinese Control Conference*, pages 3882–3886. IEEE, 2013.

[32] José Ignacio Orlando, Elena Prokofyeva, and Matthew B Blaschko. A discriminatively trained fully connected conditional random field model for blood vessel segmentation in fundus images. *IEEE transactions on Biomedical Engineering*, 64(1):16–27, 2016.

[33] Shuangling Wang, Yilong Yin, Guibao Cao, Benzheng Wei, Yuanjie Zheng, and Gongping Yang. Hierarchical retinal blood vessel segmentation based on feature and ensemble learning. *Neurocomputing*, 149:708–717, 2015.

[34] Paweł Liskowski and Krzysztof Krawiec. Segmenting retinal blood vessels with deep neural networks. *IEEE transactions on medical imaging*, 35(11):2369–2380, 2016.

[35] Huazhu Fu, Yanwu Xu, Stephen Lin, Damon Wing Kee Wong, and Jiang Liu. Deepvessel: Retinal vessel segmentation via deep learning and conditional random field. In *International conference on medical image computing and computer-assisted intervention*, pages 132–139. Springer, 2016.

[36] Nicola Strisciuglio, George Azzopardi, Mario Vento, and Nicolai Petkov. Supervised vessel delineation in retinal fundus images with the automatic selection of b-cosfire filters. *Machine Vision and Applications*, 27(8):1137–1149, 2016.

[37] Razieh Ganjee, Reza Azmi, and Behrouz Gholizadeh. An improved retinal vessel segmentation method based on high level features for pathological images. *Journal of medical systems*, 38(9):1–9, 2014.

[38] Muhammad Moazam Fraz, Paolo Remagnino, Andreas Hoppe, Bunyarit Uyyanonvara, Alicja R Rudnicka, Christopher G Owen, and Sarah A Barman. An ensemble classification-based approach applied to retinal blood vessel segmentation. *IEEE Transactions on Biomedical Engineering*, 59(9):2538–2548, 2012.

[39] György Kovács and Andras Hajdu. A self-calibrating approach for the segmentation of retinal vessels by template matching and contour reconstruction. *Medical image analysis*, 29:24–46, 2016.

[40] George Azzopardi, Nicola Strisciuglio, Mario Vento, and Nicolai Petkov. Trainable cosfire filters for vessel delineation with application to retinal images. *Medical image analysis*, 19(1):46–57, 2015.

[41] Iching Liu and Ying Sun. Recursive tracking of vascular networks in angiograms based on the detection-deletion scheme. *IEEE Transactions on medical imaging*, 12(2):334–341, 1993.

[42] Lucia Espona, María J Carreira, MG Penedo, and Marcos Ortega. Retinal vessel tree segmentation using a deformable contour model. In *2008 19th International Conference on Pattern Recognition*, pages 1–4. IEEE, 2008.

[43] Luo Gang, Opas Chutatape, and Shankar M Krishnan. Detection and measurement of retinal vessels in fundus images using amplitude modified second-order gaussian filter. *IEEE transactions on Biomedical Engineering*, 49(2):168–172, 2002.

[44] KW Sum and Paul YS Cheung. Vessel extraction under non-uniform illumination: a level set approach. *IEEE Transactions on Biomedical Engineering*, 55(1):358–360, 2007.

[45] Adrian Galdran, M Meyer, Pedro Costa, A Campilho, et al. Uncertainty-aware artery/vein classification on retinal images. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 556–560. IEEE, 2019.

[46] Meindert Niemeijer, Xiayu Xu, Alina V Dumitrescu, Priya Gupta, Bram Van Ginneken, James C Folk, and Michael D Abramoff. Automated measurement of the arteriolar-to-venular width ratio in digital color fundus photographs. *IEEE Transactions on medical imaging*, 30(11):1941–1950, 2011.

[47] Qiao Hu, Michael D Abràmoff, and Mona K Garvin. Automated separation of binary overlapping trees in low-contrast color retinal images. In *International conference on medical image computing and computer-assisted intervention*, pages 436–443. Springer, 2013.

[48] José Ignacio Orlando, Joao Barbosa Breda, Karel Van Keer, Matthew B Blaschko, Pablo J Blanco, and Carlos A Bulant. Towards a glaucoma risk index based on simulated hemodynamics from fundus images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 65–73. Springer, 2018.

[49] Behdad Dashtbozorg, Ana Maria Mendonça, and Aurélio Campilho. An automatic graph-based approach for artery/vein classification in retinal images. *IEEE Transactions on Image Processing*, 23(3):1073–1083, 2013.

[50] Samaneh Abbasi-Sureshjani, Iris Smit-Ockeloen, Jiong Zhang, and Bart Ter Haar Romeny. Biologically-inspired supervised vasculature segmentation in slo retinal fundus images. In *International Conference Image Analysis and Recognition*, pages 325–334. Springer, 2015.

[51] P Kevin Raj, Aniketh Manjunath, JR Harish Kumar, and Chandra Sekhar Seelamantula. Automatic classification of artery/vein from single wavelength fundus images. In *IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 1262–1265. IEEE, 2020.

[52] Wenao Ma, Shuang Yu, Kai Ma, Jiexiang Wang, Xinghao Ding, and Yefeng Zheng. Multi-task neural networks with spatial activation for retinal vessel segmentation and artery/vein classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 769–778. Springer, 2019.

[53] Dana Angluin. Queries and concept learning. *Machine learning*, 2(4):319–342, 1988.

[54] Eric B Baum and Kenneth Lang. Query learning can work poorly when a human oracle is used. In *International joint conference on neural networks*, volume 8, page 8, 1992.

[55] Les E Atlas, David A Cohn, and Richard E Ladner. Training connectionist networks with queries and selective sampling. In *Advances in neural information processing systems*, pages 566–573. Citeseer, 1990.

[56] David D Lewis and William A Gale. A sequential algorithm for training text classifiers. In *Special Interest Group on Information Retrieval(SIGIR)'94*, pages 3–12. Springer, 1994.

[57] Varun Chandrasekaran, Kamalika Chaudhuri, Irene Giacomelli, Somesh Jha, and Songbai Yan. Exploring connections between active learning and model extraction. In *29th USENIX Security Symposium (USENIX Security 20)*, pages 1309–1326. USENIX Association, August 2020.

[58] Burr Settles. Active learning. *Synthesis lectures on artificial intelligence and machine learning*, 6(1):1–114, 2012.

[59] Yifan Fu, Xingquan Zhu, and Bin Li. A survey on instance selection for active learning. *Knowledge and information systems*, 35(2):249–283, 2013.

[60] H Sebastian Seung, Manfred Opper, and Haim Sompolinsky. Query by committee. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 287–294, 1992.

[61] Devis Tuia, Frédéric Ratle, Fabio Pacifici, Mikhail F Kanevski, and William J Emery. Active learning methods for remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 47(7):2218–2232, 2009.

[62] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Xiaojiang Chen, and Xin Wang. A survey of deep active learning. *arXiv preprint arXiv:2009.00236*, 2020.

[63] Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. *Journal of machine learning research*, 2(Nov):45–66, 2001.

[64] Prateek Jain and Ashish Kapoor. Active learning for large multi-class problems. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 762–769. IEEE, 2009.

[65] Nabiha Asghar, Pascal Poupart, Xin Jiang, and Hang Li. Deep active learning for dialogue generation. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 78–83, Vancouver, Canada, August 2017. Association for Computational Linguistics.

[66] Tao He, Xiaoming Jin, Guiguang Ding, Lan Yi, and Chenggang Yan. Towards better uncertainty sampling: Active learning with multiple views for deep convolutional neural network. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1360–1365. IEEE, 2019.

[67] Natalia Ostapuk, Jie Yang, and Philippe Cudré-Mauroux. Activelink: deep active learning for link prediction in knowledge graphs. In *The World Wide Web Conference*, pages 1398–1408, 2019.

[68] Hiranmayi Ranganathan, Hemanth Venkateswara, Shayok Chakraborty, and Sethuraman Panchanathan. Deep active learning for image classification. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3934–3938. IEEE, 2017.

[69] Sheng-Jun Huang, Jia-Wei Zhao, and Zhao-Yang Liu. Cost-effective training of deep cnns with active model adaptation. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1580–1588, 2018.

[70] Sheng-Jun Huang, Miao Xu, Ming-Kun Xie, Masashi Sugiyama, Gang Niu, and Songcan Chen. Active feature acquisition with supervised matrix completion. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1571–1579, 2018.

[71] Hong-Min Chu and Hsuan-Tien Lin. Can active learning experience be transferred? In *IEEE 16th International Conference on Data Mining (ICDM)*, pages 841–846. IEEE, 2016.

[72] Barret Zoph and Quoc V Le. Neural architecture search with reinforcement learning. *arXiv preprint arXiv:1611.01578*, 2016.

[73] Yonatan Geifman and Ran El-Yaniv. Deep active learning with a neural architecture search. *arXiv preprint arXiv:1811.07579*, 2018.

[74] Borja Ayerdi and Manuel Graña. Random forest active learning for retinal image segmentation. In *Proceedings of the 9th International Conference on Computer Recognition Systems CORES 2015*, pages 213–221. Springer, 2016.

[75] Dwarikanath Mahapatra, Behzad Bozorgtabar, Jean-Philippe Thiran, and Mauricio 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*, pages 580–588. Springer, 2018.

[76] IS Hephzi Punithavathi, P GaneshKumar, et al. Annotation and retrieval of retinal images using support vector machine with active learning. In *IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS)*, pages 1–6. IEEE, 2019.

[77] Ruben Hemelings, Bart Elen, Joao Barbosa-Breda, Sophie Lemmens, Maarten Meire, Sayeh Pourjavan, Evelien Vandewalle, Sara Van de Veire, Matthew B Blaschko, Patrick De Boever, et al. Accurate prediction of glaucoma from colour fundus images with a convolutional neural network that relies on active and transfer learning. *Acta ophthalmologica*, 98(1):e94–e100, 2020.

[78] Wei Li, Mingquan Zhang, and Dali Chen. Fundus retinal blood vessel segmentation based on active learning. In *International Conference on Computer Information and Big Data Applications (CIBDA)*, pages 264–268. IEEE, 2020.

[79] Firat Ozdemir, Zixuan Peng, Philipp Fuernstahl, Christine Tanner, and Orcun Goksel. Active learning for segmentation based on bayesian sample queries. *Knowledge-Based Systems*, 214:106531, 2021.

[80] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[81] Shibani Santurkar, Dimitris Tsipras, Andrew Ilyas, and Aleksander Mądry. How does batch normalization help optimization? In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 2488–2498, Red Hook, NY, USA, 2018. Curran Associates Inc.

[82] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[83] Xinyang Deng, Qi Liu, Yong Deng, and Sankaran Mahadevan. An improved method to construct basic probability assignment based on the confusion matrix for classification problem. *Information Sciences*, 340-341:250–261, 2016.

[84] Carole H Sudre, Wenqi Li, Tom Vercauteren, Sebastien 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*, pages 240–248. Springer, 2017.

[85] Xiayu Xu, Rendong Wang, Peilin Lv, Bin Gao, Chan Li, Zhiqiang Tian, Tao Tan, and Feng Xu. Simultaneous arteriole and venule segmentation with domain-specific loss function on a new public database. *Biomed. Opt. Express*, 9(7):3153–3166, Jul 2018.

[86] Liangzhi Li, Manisha Verma, Yuta Nakashima, Ryo Kawasaki, and Hajime Nagahara. Joint learning of vessel segmentation and artery/vein classification with post-processing. In *Medical Imaging with Deep Learning*, 2020.