Detection of Object with Less Visual Information

PHO NGOC DANG KHOA

Japan Advanced Institute of Science and Technology

Doctoral Dissertation

Detection of Object with Less Visual Information

PHO NGOC DANG KHOA

Supervisor : Atsuo YOSHITAKA

Graduate School of Advanced Science and Technology

Japan Advanced Institute of Science and Technology

Information Science

September, 2022

Abstract

Object Detection plays an essential role in many practical applications such as video analysis, image understanding, security, etc. Recent years have witnessed the breakthroughs of Deep Learning based methods in various applications in daily life images. Deep Learning-based methods achieve human-level performances on classification tasks for identifying successfully more than tens of thousands of categories such as animal species, vehicles, household objects, etc. However, deep neural networks are still limited in specific medical image domains. The neural networks often fail for reasons such as a small number of training data, difficulty characterizing the target objects, or a small number of differences among the target objects.

This study is motivated by the need for a comprehensive method for medical image domains to accelerate the diagnosis and treatment processes. However, the target objects of medical image domains often (i) have less visual information, (ii) have a small number of training data, and (iii) have various appearances. This dissertation detects objects with a few features appearing in medical images. This study assumes that each target category has enough distinctive features to identify even with a small number of features. Finding those distinctive features is essential to detecting objects with less visual information.

In the case of small training data, the problem becomes more difficult since the method must find the correct distinctive features within a few samples. A particular surrounding object may have a high probability of appearing along with the target object. With a small number of training samples, the deep neural network detectors view the background information to identify the objects. This study focuses on finding features that characterize the target objects rather than the unique background features to overcome the problem. The segmentation-driven mechanism is proposed to guide the detector to focus only on the regions of the target objects. The mechanism is integrated into a neural

network detector to form Segmentation-driven RetinaNet to filter out the background by the segmentation mask and then detect and identify the objects in the filtered image.

While the characteristic features are efficient for detection tasks, distinctive features are essential for identifying the objects. Detecting objects in grayscale images is also a challenging problem. Grayscale images with only one color channel have much smaller feature spaces than general color images. The objects in grayscale images are characterized only by outer shapes, connectivity, and the intensity of the pixels. The attention-driven mechanism is proposed by replacing the segmentation with the attention mask to guide the deep network in focussing on the distinctive features of the target objects.

Finally, this study explores the relationship among the various appearances of a target object category. An object category may have several appearances, and each appearance only shows some category features. In many cases, an appearance of one category may have more similarities to that of another category than its intraclass appearances. Motivated by the taxonomy of animals, this study investigates the hierarchical multi-label classifier and the category hierarchy structure. Training samples of each category are clustered concerning the appearances. Multiple labels following the category hierarchy structure are assigned to training samples. The hierarchical classifier is integrated with the Segmentation-driven RetinaNet to form a unified network for detection.

Experiments are conducted on realistic datasets from the protozoa and DNA Profiling domains as examples of objects with less visual information in color and grayscale images, respectively. Experiments show that the Attention-driven mechanism effectively guides the neural network detectors to find the distinctive characteristic features of the target objects. Even with at most five samples per subcategory for training, this study successfully trained the proposed method for detecting the protozoa in the micrographs. With 16 training samples, the proposed method achieves the highest performance on the DNA Profiling image dataset. Besides, the integrated hierarchical multi-label classifier boosts the detection performance for the polymorphism problems.

Keywords: Detection, **Segmentation**, **Identification**, **Protozoa**, **Genome Profiling**

Acknowledgments

First, I would like to thank my supervisor, Assoc. Prof. Atsuo Yoshitaka of the School of Information, Japan Advanced Institute of Science and Technology. I would like to express my sincere gratitude to the support of my study, for his patience, motivation, and enthusiasm. Without his guidance, I wouldn't finish my dissertation.

Second, I would also like to thank Assoc. Prof. Bac Le for his guidance on my research career. I would like to express my sincere gratitude for giving me chance of the life time to come to JAIST.

Special thanks to all the members of Yoshitaka Laboratory. Thank you guys for sharing the funs throughout the time. I would like to thank Anh-san for the discussion to help me find the idea for this dissertation.

Finally, I must express my deepest gratitude to my parents for providing me with unfailing support and continuous encouragement throughout my years of study.

Ishikawa, Japan

Pho Ngoc Dang Khoa

Contents

List of Figures

List of Tables

Terms and Abbreviations

FPN : Feature Pyramid Network

mAP : Mean average precision

TGGE : Temperature gradient gel electrophoresis

spiddos : Species Identification Dots

Chapter 1

Introduction

1.1 Background

This dissertation focuses on the detection, segmentation, and identification tasks of objects in 2D images. The identification task aims to predict the class or category of the object that appears in the input image. The identification tasks assumes that the input image contains exactly one object instance belonging to one of the predefined target categories to make it easier. The object takes the majority of the area of the input image. Localization aims to predict the position of the object instance in the input image. In localization, the input images also are assumed to contain exactly one object. Unlike identification problems, the object takes a smaller proportion in the images. The problem would be more complex and challenging for typical images that contain more than one object. The detection aims to localize and identify the multiple objects in the input image. Detection is more complex than identification and localization since the detection problems require the systems to indicate multiple objects or perceive when no object is in the images. The segmentation aims to indicate which pixels belong to the object instances. Figure 1.1 shows an example of detection, segmentation, and identification tasks.

Object Detection plays an essential role in Computer Vision. Recent years have witnessed a breakthrough in Deep Learning-based methods in various applications in daily images. Those methods are able to exceed human performances on large-scale image

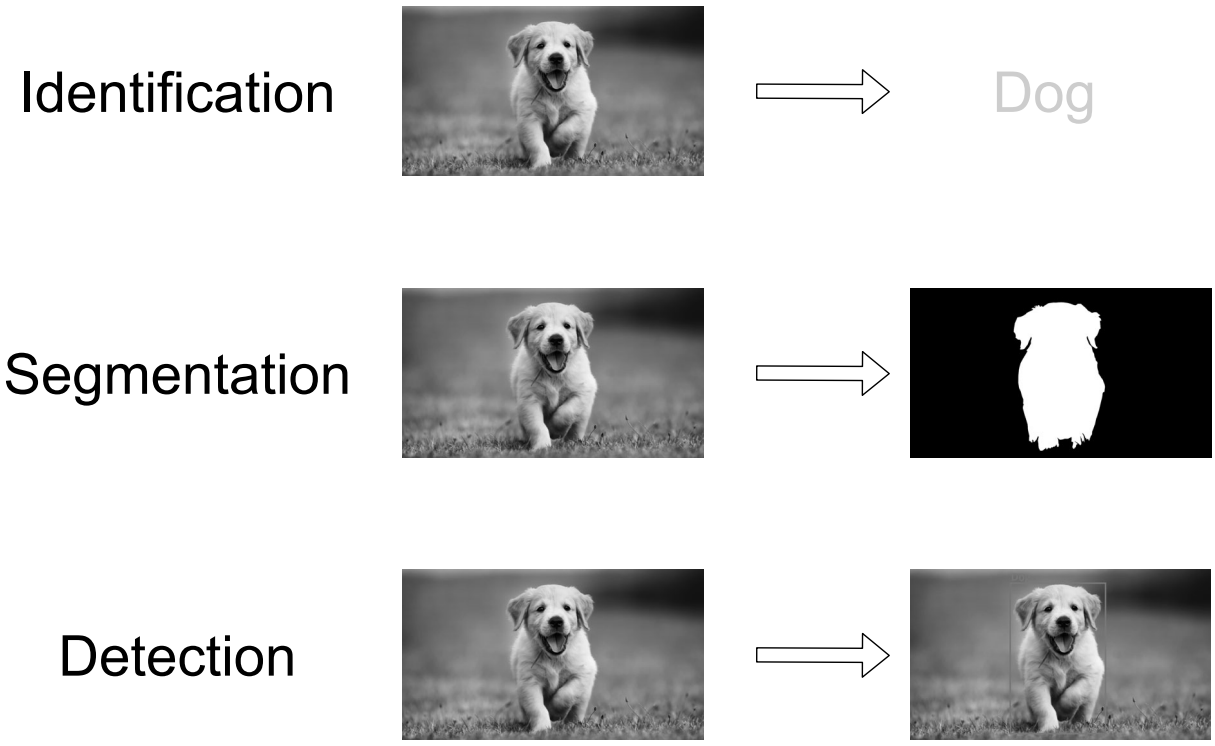


Figure 1.1: Examples of identification, segmentation, and detection task. (Image credit: Shutterstock)

datasets which contain thousands of categories such as ImageNet [5]. However, in specific domains such as medical image analysis, the Deep Learning-based methods often fail for several reasons, such as the small number of training data [6; 7], or a small number of differences among the target objects [8].

Since Deep Learning requires a massive amount of training datasets, they do not succeed on small datasets or the domains that are difficult to collect training data. Various appearances of a category make it difficult to capture the generalized appearance. Very similar objects require the method's capabilities to find the most distinctive features among those objects.

The unavailability of the dataset is one of the critical problems of deep learning in medical imaging [6]. The development of large medical imaging data is quite challenging since the annotation from the experts requires extensive time and often requires multiple expert opinions to overcome human error. While Deep Learning-based detection methods require many training data samples to achieve good performances, detecting target objects in small training sets remains a problem. The input image contains an arbitrary number of

target object instances. Therefore, each region in the input image is treated independently. In an image, the number of regions containing the target object instances is much less than the number of background regions. It leads to the class imbalance problem from the target categories and background. The small number of training data makes it more difficult to train a detector for the target categories.

This dissertation focuses on detecting objects with less visual information in specific domains such as medical image domains. They are objects that have fewer distinctive features compared to ordinary objects. This study aims to learn the most distinctive features among various appearances and similar categories of objects with less visual information. This study chooses protozoa micrograph and DNA profiling images as examples of objects with less visual information. Figure 1.2 shows an example of protozoa, whose target species are Giardia lamblia (Gla), Iodamoeba butschilii (Ibu), Toxoplasma gondi (Tgo), Cyclospora cayetanensis (Cca), Balantidium coli (Bco), Sarcocystis (Sar), Cystoisospora belli (Cbe) and Acanthamoeba (Aca). Figure 1.3 shows an example of DNA profiling image.

In general, an object instance can be decomposed into multiple components. Each component can be decomposed into multiple parts. The decomposing process can be applied further until reaching lines and curves. The relationship of decomposed parts can be captured as a tree structure. Figure 1.4 shows some examples of part-based tree structures of general objects such as planes and trucks. The depth of the tree is $log(p)$, where p is the number of primitive parts that can be decomposed from the object.

This study focuses on detecting and identifying objects with less visual information. Compared with other general objects, the objects with less visual information have much shallower part-level hierarchical tree structures. This study selects the domain of protozoa micrograph as an example for objects in color images and the DNA profiling image domain as examples for objects in grayscale images. Figure 1.5 shows the structure of some protozoa in biology field. However, the appearances of protozoa show only the cell wall, nucleus, and sometimes flagellum, while other parts are invisible in the micrograph. The depth of protozoa part-based hierarchical tree structure is around 1 or 2. Figure 1.6

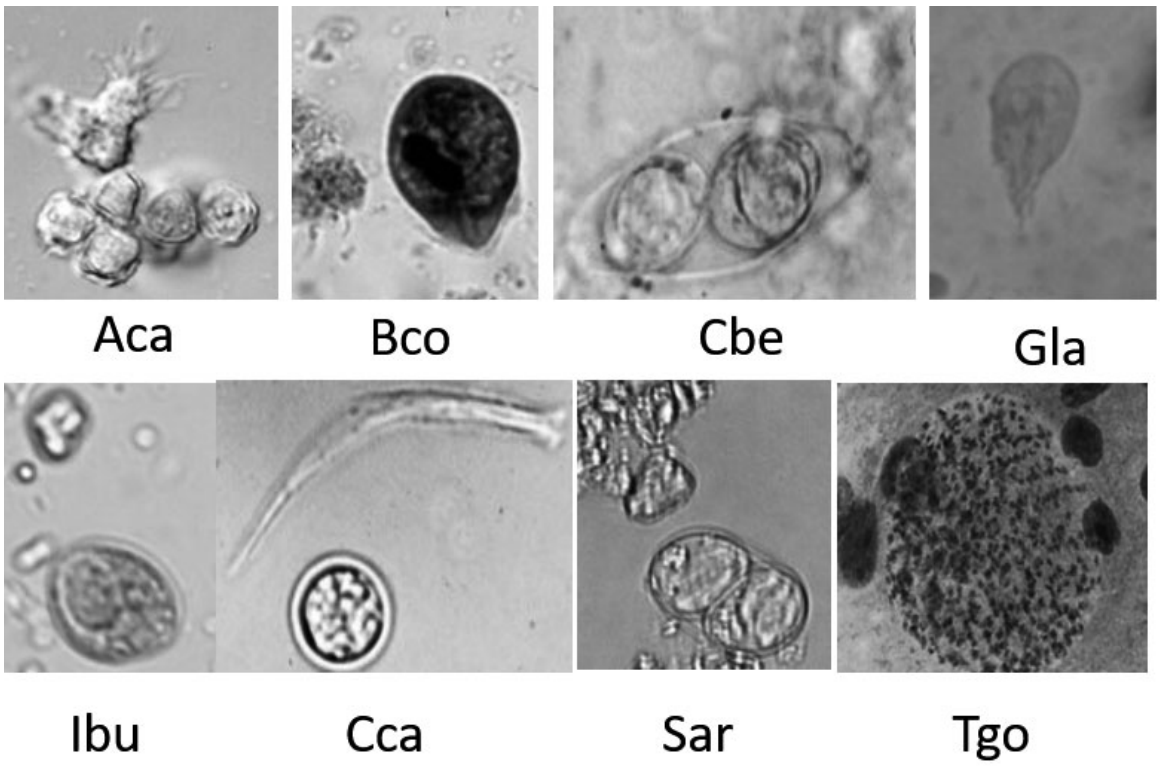
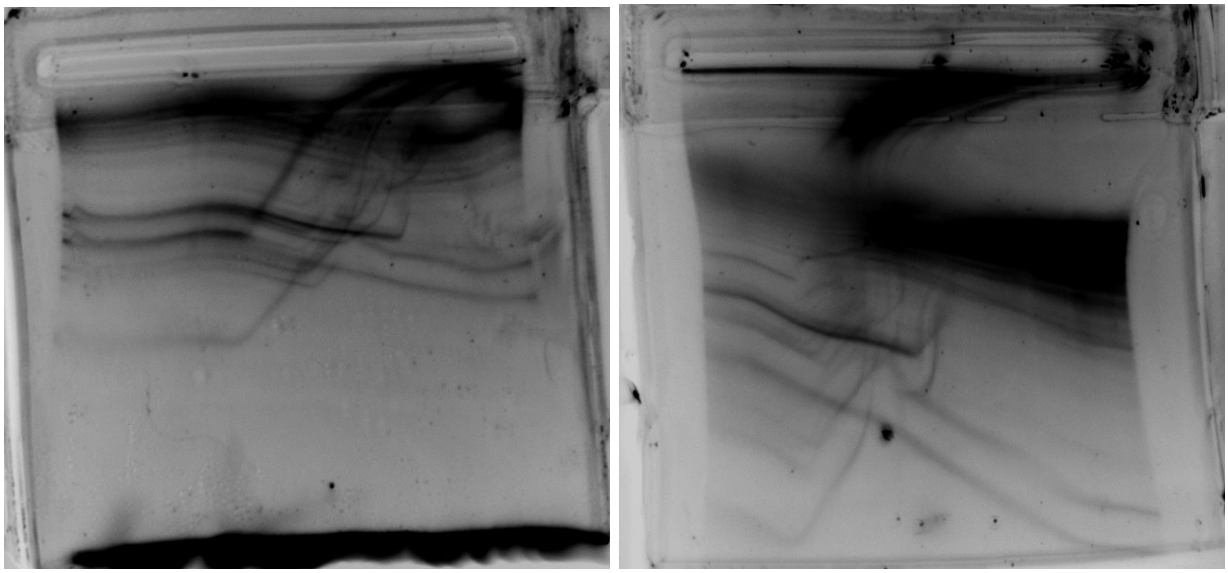


Figure 1.2: Examples of the target protozoa. (The micrographs were provided by Dr. Masaharu Tokoro at Kanazawa University)



(a) (b)

Figure 1.3: An example of DNA Profiling image. (The images were provided by Professor Kiyoshi Yasukawa at Kyoto University)

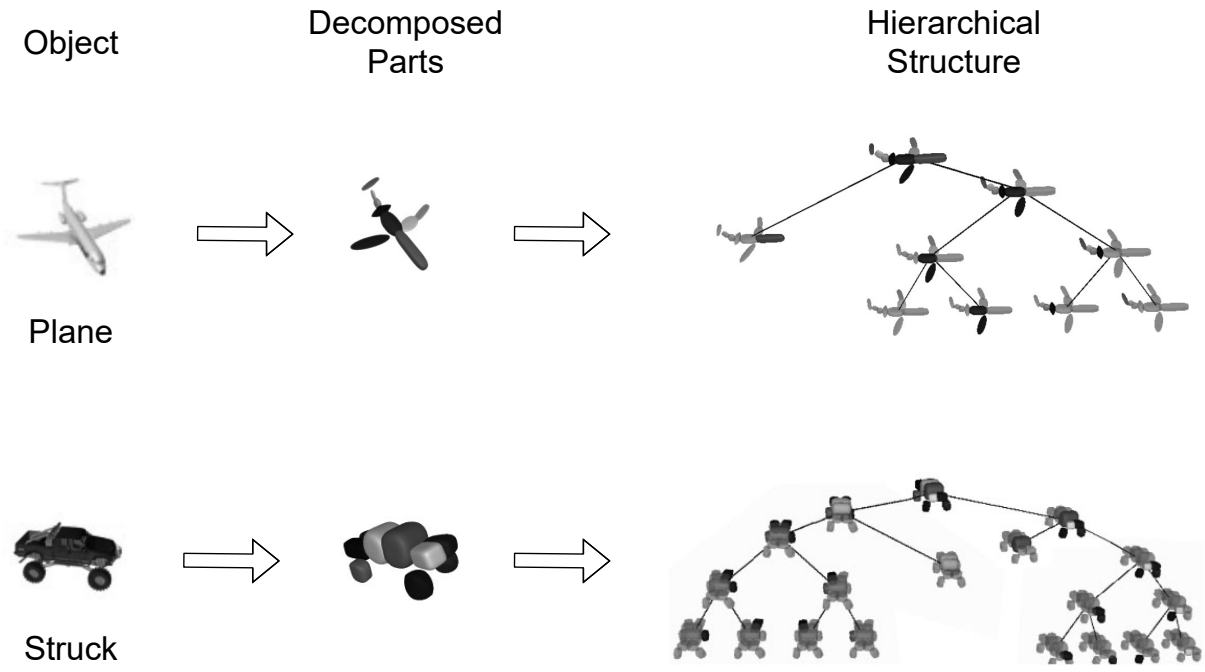


Figure 1.4: Examples of part-based hierarchical tree structure. (Image taken from [9])

shows an example of part hierarchy of protozoa. In the example, the instance of the protozoa can be decomposed into cell wall, macronucleus, and micronucleus. The other parts are invisible in this image. It is difficult to decompose those parts further. DNA profiling images contain trajectories in grayscale. The trajectories contain lines, curves, and differences in pixel intensities. The part-based hierarchical tree structure is also shallower than other general objects. Therefore, protozoa micrograph and DNA profiling images can be considered objects with less visual information.

1.2 Research Motivation

Automated analysis methods for medical image domains are in demand. It would reduce the time and human effort to diagnose and give treatment. Moreover, training medical experts also require a lot of time and cost. Therefore, automated methods are significant when analyzing, diagnosing, and treating large groups of people, such as crises, disasters, or pandemics. However, there is no "silver bullet" for all the existing problems. This study is motivated by the need for a comprehensive method for medical image domains

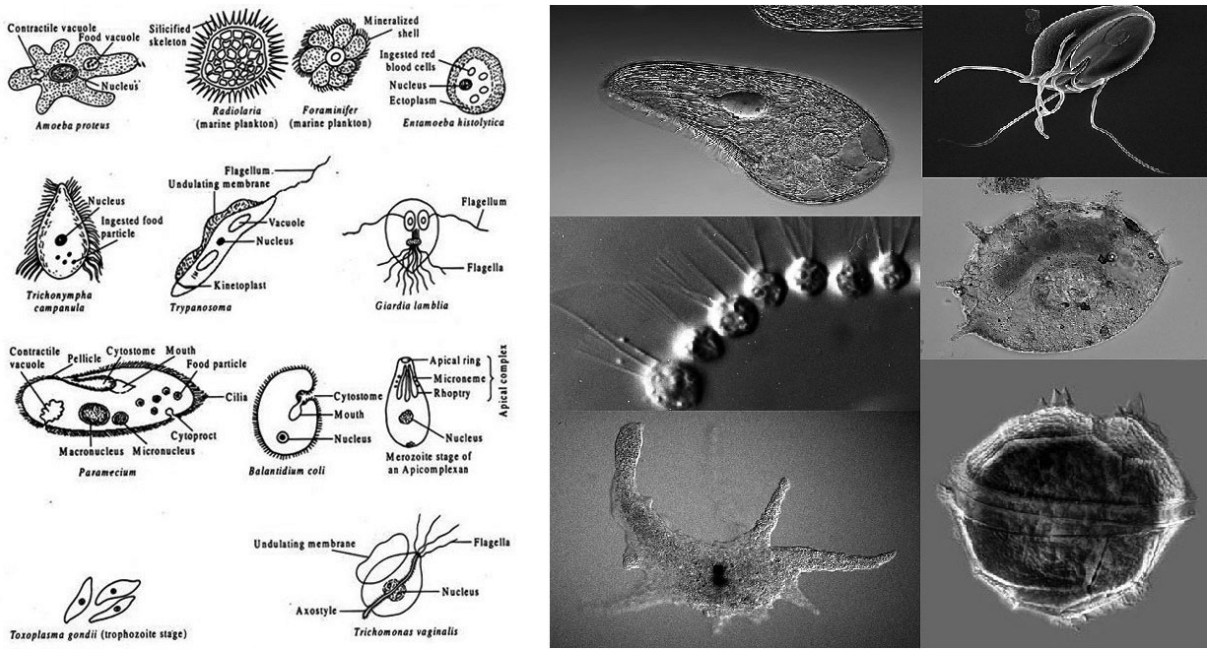


Figure 1.5: Detail parts of protozoa in biology field. (Image credit: biologydiscussion.com)

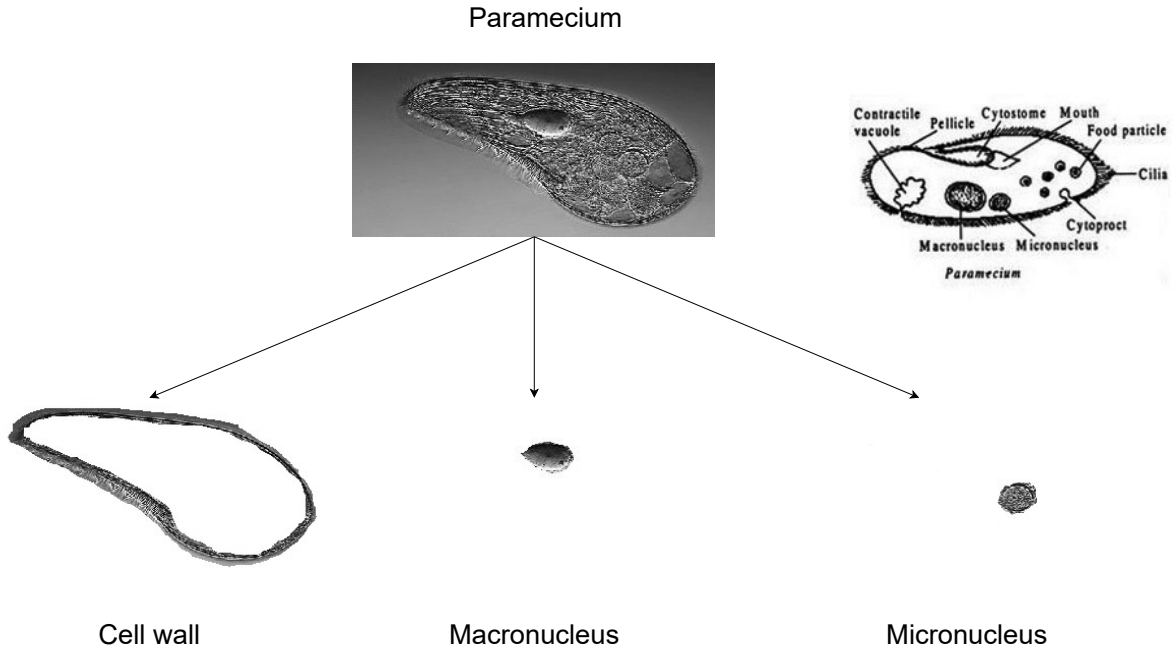


Figure 1.6: Example of part hierarchy of protozoa

to accelerate the diagnosis and treatment processes.

Building an automated system for medical image domains requires computer vision techniques and medical knowledge. The target objects in medical domains are often captured by special devices. The characteristics of target objects are therefore different compared to daily life objects. The visual features of those objects is not easy to formalize. That leads to the challenge of learning to generalize those medical objects with the computer.

What will be elucidated in this dissertation:

- Is it feasible to detect and identify objects with a small amount of information?
- What kind of features should be detected from the target objects?

Those are the underlying fundamental questions to pursue this dissertation. An object has its original features that characterize its appearance. However, its appearance changes to adapt to the changes in the environment. The original features and their changes make the object unique even though, in most cases, the original features are not shown in the images. Capturing the characterizing features and the unique changes is the key to detecting objects with less visual information.

1.3 Research Question and Solutions

The dissertation's research question is: "How can we identify a category with a small number of visual information?". An object can be described by its characteristic features. Based on the characteristic features, the appearance of the object in the image can be reconstructed. However, the characteristic feature sets of two different categories may overlap. Generally, the structures of target objects change to adapt to the environment during evolution and natural selection. By external influences, the appearance may also roughly change during the lifetime. This dissertation assumes that even though the appearance roughly changes, each target category has distinctive features that are enough to identify. This assumption holds for the case of an object with less visual informa-

tion. Finding those distinctive features is the key to detecting objects with less visual information.

The Deep Learning model will learn the characteristics or distinctive features depending on the purpose. The characteristic features will be learned in applications such as image generating or image synthesis. From the learned features, the generator produces an image toward the goal of the task. On the other hand, image classification learns distinctive features to distinguish the differences between the target objects. Generally, learning all the distinctive features is unnecessary. A few distinctive features are may enough to classify the category among the target categories. This research aims to find the characteristic features that are also distinctive to characterize the target objects. The characteristic features are represented as latent factors or latent structural relationships that make the target objects different from each other. To answer the research question, this research focuses on the following problems:

- **Finding characteristic features**: This study focuses on finding characteristic features of the target objects in the case of the small dataset and the case of grayscale images. The interesting characteristic features are shape features and texture features. In the case of small datasets, irrelevant features or backgrounds are often selected as distinctive features to detect and identify the target object. This study considers the protozoa and genome profiling images to find the texture and shape features. The segmentation-driven mechanism is proposed to guide the deep learning-based methods to focus only on the regions of the target objects.

- **Finding distinctive features**: This study is interested in the protozoa species that share similar round shapes. Textures inside the cell wall are essential to identify those protozoa. In genome profiling images, objects are characterized only by the outer shape, the connectivity, and the intensities of the pixels. The primary to improve the identification accuracy of the target objects in both cases is to find the distinctive features. The attention-driven mechanism is proposed to refine the segmentation results and focus on the distinctive features.

- **Relationship of appearances in the polymorphism of the appearances**: Dividing the training samples into sub-categories helps improve the identification performances in the case of polymorphism of the appearance of the target objects. It would clarify the differences between samples of a category. Since similar samples are clustered into the same sub-category, it is easier to generalize their common features. Compared to the goal of detecting categories, detecting the sub-category forces the network to learn more efficient features to distinguish the similar appearances of a different species. Dividing the training samples into sub-categories helps to find the distinctive features to distinguish various appearances of a target object. The divided sub-categories are treated independently to find the most distinctive features. However, the similarities between intraclass sub-categories, which are divided from the same category, are not considered in this manner. The unique differences contribute the most to identifying a category among the target category set. To reveal the similarities of intraclass sub-categories, hierarchical multi-label classifiers (in place of flat multiclass classifiers) are applied to this problem. Hierarchical multi-label models provide a better view of prediction stages, including taxonomy predictions. A hierarchical neural network classifier is proposed to solve this problem.

Figure 1.7 shows the main structure of the sub-topics of this dissertation to solve this research question. Each input sample only contains a smaller set of the characteristic feature set of the target category. To identify the category of the input sample, a set of characteristic features must be extracted correctly. This set of features is often smaller than the distinctive feature set of the category. The task is to determine which category the smaller set belongs to. Chapter 3 and 4 aim to extract the characteristic and distinctive features from the input images. Chapter 5 explores the relationship between appearances of the same category.

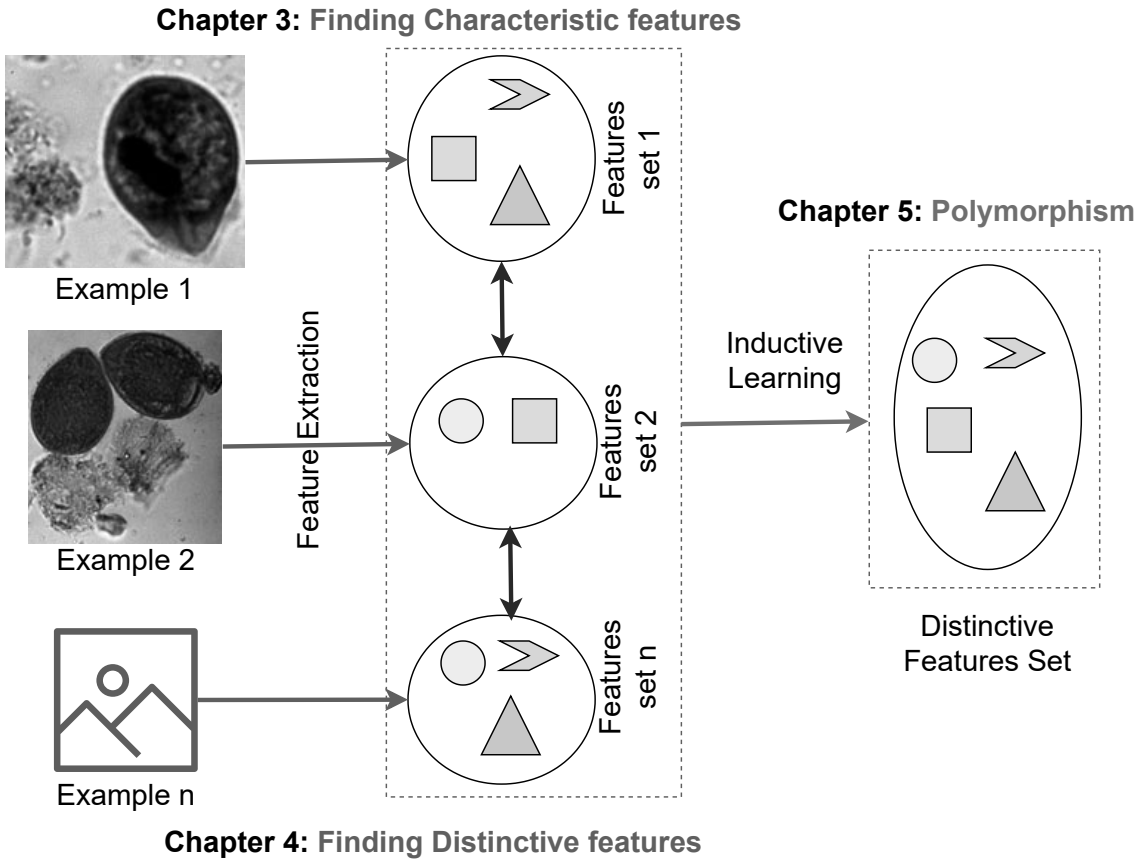


Figure 1.7: Topic structure of this dissertation.

1.4 Technical Challenges

Object Detection is an important and challenging research topic. There are several challenges in the detection of objects with less visual information, such as:

- **Finding the characteristic features among similar object category**: Finding characteristic features is essential to detecting the target objects. To annotate objects in the images, the most common method is to draw rectangle bounding boxes around the objects. Since most annotators are only specialized for a specific domain, annotations which are more complex than rectangle bounding boxes are not affordable. The target objects do not often take entire regions indicated by the bounding boxes, irrelevant objects or backgrounds are included in the annotations. In small datasets, a specific irrelevant feature may have a high probability of appearing along with the target category. The problem becomes more sensitive when similar categories are in the target category set. Since there are few differences, the

irrelevant features have high probabilities of being chosen as distinctive features.

- **Small number of training data**: In general, objects with less visual information don't often come with a small number of training data problems. With enough training data, the Deep Learning-based methods can learn a category's common features and the most distinctive features to distinguish it from other categories. Detecting objects with less visual information is also achieved with enough number of training data. However, in the case of small training data, the problem becomes more difficult since the method is required to be adapted the correct distinctive features within a few samples. Almost all the category features appear one time through the small data; thus those features can be chosen as distinctive features to distinguish from other categories. The methods are easy to fall into learning irrelevant features; therefore, they often fail to detect unseen object's appearances in the test images. This study focuses on detecting objects with less visual information in the case of the small number of training data.

- **Polymorphism of the appearances of the target**: The objects may have various appearances in different images. For example, an animal grows over time; thus, the image's size changes. Images of a face taken from different poses and also different appearances. In this study, this phenomenon is refered to polymorphism of appearance. The polymorphism of appearance makes it difficult to capture the generalized features of the category. A deep understanding of the causes of polymorphism of appearance helps to obtain the characteristic features and eliminate the irrelevant features. The target objects show different appearances in different image samples. Each appearance only shows some of the characteristic features of the category. To correctly identify the category, the characteristic and distinctive features in the appearance should be detected to infer the category's feature set.

1.5 Contribution

This dissertation detects and identifies objects with less visual information in the images. To that end, this study aims to (1) find the characteristic features of the target objects, (2) find the distinctive features of the target objects and (3) explore the relationship between the appearances of the target objects. Experimental evaluations are performed on various datasets, including the protozoa dataset and DNA Profiling image dataset.

The contributions of this dissertation are as follows:

- **Establishing Segmentation-driven mechanism** to find the characteristic feature of the target objects. The proposed mechanism guides the Deep Learning-based methods to focus on the regions of the target objects. In the case of the small training data, this dissertation introduces Segmentation-driven RetinaNet to detect, segment, and identify the target objects. The proposed method applied segmentation to filter out all the background, then detect and identify the object. Experiments are conducted in the protozoa dataset to show the effectiveness the the proposed method. Even with at most five samples per sub-category for training, the proposed method can successfully detect and identify the species of protozoa.

- **Establishing Attention-driven mechanism** to find the distinctive feature of the target objects. The attention mechanism guides the Deep Learning-based methods to focus on the distinctive features that contribute for identification. This dissertation introduces the Attention-driven RetinaNet for both protozoa and DNA profiling domains. The segmentation mask is replaced with the attention mask, which focuses on the important parts of the object instances in the input image. Even with 16 images for training for the DNA profiling domain, the proposed method achieves promising results for clinical analysis.

- **Integrating hierarchical multi-label classifier** for detecting the objects with less visual information. This study explores the relationship of appearances of the target categories to solve the polymorphism of the appearance problem. Charac-

terizing the target objects based on their hierarchical relationships improves the learning process on common features of categories and distinctive features of subcategories. This study introduces Segmentation-driven Hierarchical RetinaNet that integrates the hierarchical classifier with a detection network for the less visual information objects. Applying the hierarchical structure for clustering the different appearances clears the ambiguity in each category. Experiments are conducted on the protozoa dataset. In these experiments, 5 samples per life-cycle stage are being used to train the detection methods. The proposed method achieves the highest mAP, precision, and recall values over the related works.

1.6 Dissertation Organization

Structure of this dissertation is as follows:

- **Chapter 1: Introduction**

  This chapter mainly introduces preliminary concepts, technical challenges, research question, contribution, and the outline of this dissertation. Brief views of the proposed methods are also given.

- **Chapter 2: Related Work**

  This chapter describes a literature review on related studies on the protozoa field and DNA profiling domain. Related deep learning architectures are also reviewed.

- **Chapter 3: Finding the characteristic features**

  This chapter presents the Segmentation-driven mechanism for neural network detectors. The proposed mechanism helps the neural network to focus on finding the characteristic features of the target categories, even in the case of small datasets. The organization of the Segmentation-driven RetinaNet, which integrates the segmentation-driven mechanism into a neural network detector, is also presented in this chapter. The Segmentation-driven RetinaNet is evaluated on the protozoa dataset provided by Dr. Masaharu Tokoro at Kanazawa University.

- **Chapter 4: Finding the distinctive features**

  This chapter introduces the Attention-driven mechanism for improving the identification accuracy of neural network detectors. This mechanism focuses on finding the target objects' distinctive features and essential parts. The mechanism is evaluated on both protozoa and DNA profiling image datasets.

- **Chapter 5: Polymorphism**

  This chapter explores the relationship of various appearances of the target categories. A hierarchical classifier is proposed to be intergrated into a neural network detector to solve the polymorphism problem. The method is evaluated on the protozoa dataset.

- **Chapter 6: Conclusions and Future Work**

  Finally, a summary of the presented methods is shown in Chapter 6. Besides, this dissertation also discusses possible improvements and extensions for detecting objects with less visual information.

Chapter 2

Related Work

This section reviews existing methods for various species of protozoa and the related deep network architectures to the proposed model.

2.1 Protozoa Domain

Due to the limitation of data accessing, there are few works on the same set of species. Previous works also try to classify protozoa along with other kingdoms like metazoa or fungi. Table 2.1 shows the target species of protozoa in previous works.

To the best of our knowledge, this study is the first to perform automatic detection for protozoa images with backgrounds. Previous works only perform protozoa classification and localization where segmentation of target objects are performed manually. On the other hand, detection tasks require the capability of prediction when there are multiple instances or no instance in the image.

Most of previous works [14; 15] on segmentation are semi-automatic. They use shape features to segment out rounded shape objects and need manual selections to get the right regions of protozoa. Automatic segmentation methods try to find elliptical objects in the images [10; 16].

Suzuki et al. [10] applied ellipticity to segment out the protozoa. The ellipticity, ratio between geodesic distances, curvature variance, salience variance, red texture, average in

Table 2.1: Target Species of Previous Works

	[10]	[11]	[12]	[13]	This study
Plasmodium malariae			O		
Balantidium coli				O	
Giardia intestinalis	O	O			O
Entamoeba histolytica	O				
Cyclospora cayetanensis					O
Iodamoeba butschlii	O				O
Emdolimax nana	O				
Blastocystis spp.					
Acanthamoeba					O
Sarcocystis					O
Toxoplasma gondi					O
Cystoisospora belli					O
Trypanosoma spp.				O	

red channel, average number of regional minima in gradient, perimeter, area, and moments are extracted. They used Optimum-Path Forest [17] as the classifier to identify the species and achieved 97.46% of accuracy on protozoa. Yang et al. [18] focused on 7 parasitic eggs from helminths and applied Artificial Neural Network on size, shape and smoothness features. They tested 87 ordinary images and 100 impurity images and achieved 84% of accuracy. Flores et al. [19] extracted color density histogram and orientation information from parasite eggs in 0°, 45°, 90° and 135°. They achieved 92.16% of accuracy by applying Multiclass Support Vector Machine. Zou et al. [20] build a retrieval system by decomposing the input images into multiple color channels and extracting SIFT [21] features. Sergey Kosov et al. [22] used DeepLab-VGG-16 [23] and Conditional Random Field [24] to identify 20 species of EM.

2.2 Genome Profiling Overview

Analyzing gene information is essential for understanding the genetic changes under the effect of chemicals or the natural environment. Genome Profiling (GP) [25] is a wide-range-applicable method that has the potential for analyzing the gene information of

individual cells or species. GP helps to understand why people react differently to the same medicine and to develop proper diagnoses and treatments [26]. In the field of forensic medicine, GP can also be performed to analyze molecular phylogenetics of plasmid samples to identify the virus that infected a body when the autopsy is difficult [27; 28]. Species identification and classification are essential in many practical domains such as parasitology, microbes, biological treatment, and scientific research. The identification of species depends on the kind of species that may require different methods. For example, the Trichosporon species [29] have been classified depending on their morphological and biochemical properties, the shape of segregating cells, and the amount of xylose contained in a cell. To identify whether the blood samples are from humans, serological methods [30; 31] and DNA analysis using polymerase chain reaction (PCR) amplification. Immunological methods apply non-specific reactions for orangutans and gorillas [32; 33]. It is necessary to prepare specific antibodies of the species for immunological examinations. The human-specific sequences are detected when applying DNA analysis in the D17Z1 myoglobin gene [34; 35], D-loop region [36; 37], cytochrome b gene [38], and 16S rRNA gene of mitochondrial DNA [39], 285 rRNA [40], and TP53 gene [41]. Those conventional DNA analyses require extensive human effort, technical skill, and expensive equipment and reagents. Therefore, a simple, non-radioactive, non-toxic, and wide-range-applicable method that can be applied to any species such as GP is in demand.

In the GP method, DNA is PCR-amplified using a random primer (random PCR). Temperature gradient gel electrophoresis (TGGE) is performed. TGGE, which is reliable, reproducible, and rapid, allows the simultaneous analysis of multiple samples. TGGE is suitable for detecting known and unknown mutations in large genes where high sensitivity is required. TGGE separate DNA fragments of the same length but with different sequences. The principle of the TGGE is that their partial denaturation reduces the electrophoretic mobility of double-stranded DNA fragments. Domains that have identical melting temperatures T_m lead to the melting of DNA fragments within discrete domains. Once the domain reaches its melting temperature T_m at a particular position in the temperature gradient gel, the segment of the DNA double helix transitions turns to melted

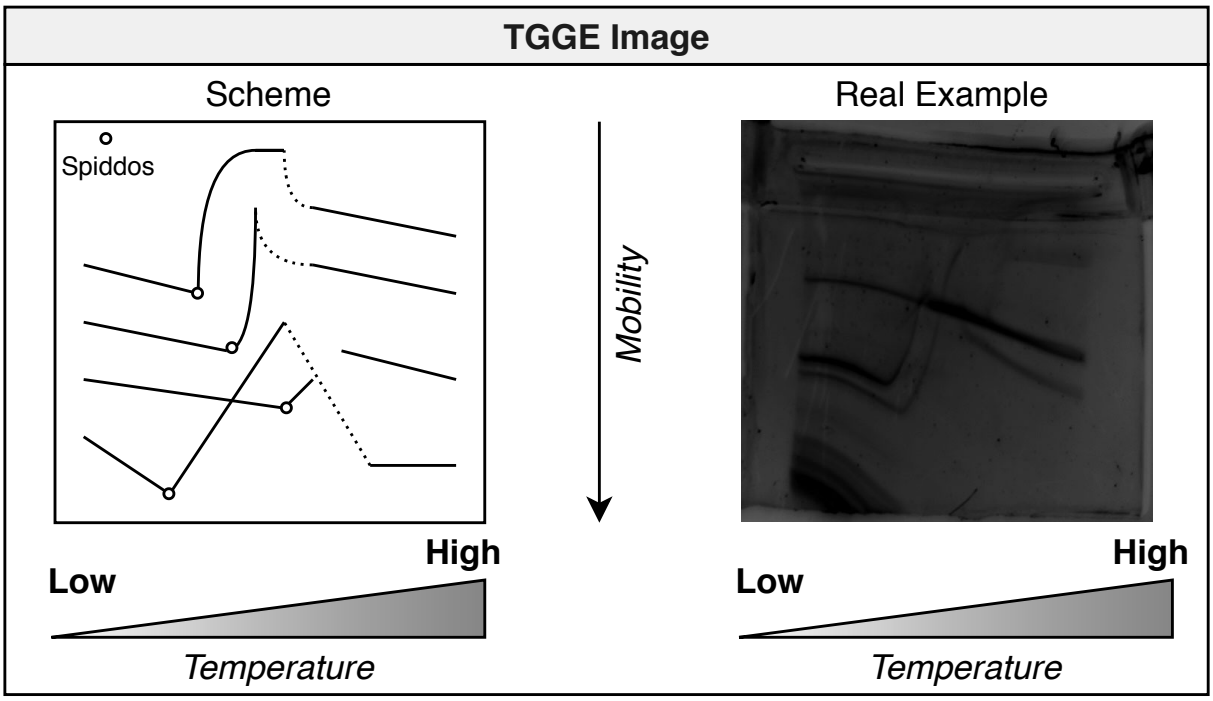


Figure 2.1: Example of TGGE images.

single-stranded DNA. The migration of the DNA molecule then stops. The segment of DNA melts from the lowest T_m to the highest T_m. Molecules with different sequences stop migrating in the temperature gradient gel at different positions. DNA fragments are now separated. The technique provides the profile of the genetic diversity of the dominant populations. DNA fragments in TGGE profiles are visualized as images of trajectories. A set of spiddos (species identification dots) that are the flexion points of the trajectories on the TGGE profile image are extracted. The spiddos pattern corresponds to the melting temperatures of the double-stranded DNA and the amplified fragments. Two types of reference DNA, whose melting temperatures are known, are prepared as TGGE internal standards to calibrate the extracted spiddos. The spiddos pattern is then compared to the referenced patterns in the database to identify the species. Pattern Similarity Score (PaSS) is calculated as a similarity metric to compare two spiddos patterns [42] for gene identification.

However, the TGGE method has limitations, such as its labor-intensive nature and intensity detection after electrophoretic separation. The image processing method is one of the keys to speeding up the species identification process. In this study, an image-based

method is proposed to extract the spiddos pattern automatically. The method takes TGGE images as input, performs preprocessing techniques to enhance the trajectories, and detects the spiddos. The TGGE image contains target trajectories that correspond to the primers. Each target trajectory has a flexion point that indicates the melting temperature of the primer. Due to the capture devices, the TGGE images also contain several uninterested lines and dots, which are considered as outliers due to the capture devices. From the image processing point of view, spiddos detection on TGGE images has several technical challenges, such as the target trajectories being disconnected, heavy noise, unstable brightness, trajectory overridden, and similar appearances between the spiddos and uninterested flexion points. It is difficult to keep track of the disconnected lines. Ridge detection [43; 44; 45] is able to detect the line in the form of multiple small parts; that is hard to keep track of the line and find out the coordinate of the spiddos. Heavy noise, unstable brightness, and trajectory overridden prevent global image processing methods from enhancing the target trajectories.

The Genome Profiling method is developed by Nishigaki et al. [25] in the bio-industry field and has been applied for many practical domains. The GP method can be applied to identify species, including human [46; 47; 48], plants [49; 50; 51], bacteria [52; 53; 54], insect [55; 56], and fungi [29]. Kinebuchi et al. [27] applied GP to identify human and other 12 species that may be found at crime case. Thanakiakrai et al. [57] identifies the meat species by using low resolution melting for two direct-triplex real-time PCR. [58; 59] clusters the species and draw the taxonomy by grouping the closest genomes of the species. However, GP requires a large amount of labor to perform and intensity detection after electrophoretic separation. The domain experts decide the number of DNA fragments in the test and the references with its certain melting temperatures. Spiddos patterns are assigned manually by the domain experts. This study aims to build an image processing method that can automatically extract the spiddos patterns from the TGGE images.

2.3 Neural Network Architecture

Object Detection has a long and rich history in the Computer Vision field. One of the earliest successes is the classic object detector with the sliding-window paradigm. Lecul et al. [60; 61] apply convolutional neural networks for handwritten digits. Viola and Jones [62] apply boosted cascade object detector for face detection. Felzenszwalb et al. proposed Deformable Part-based Model [63; 64; 65] to detect objects and its parts in the images.

With the resurgence of deep learning, one-stage and two-stage detectors came to dominate for Object Detection. In two-stage detectors, the first stage produces a sparse set of candidate proposals that may contain all the objects, then the second stage classifies the proposals into whether foreground or background classes. One-stage detectors directly predict object bounding boxes for images with no intermediate task needed. Modern Deep learning-based detectors apply backbone network as the basic feature extractor toward different requirements on accuracy and efficiency. Backbone networks for detection are networks for classification tasks taking out the last fully connected layer. Deeper and densely connected backbone such as Resnet [66], ResNeXt [67], or AmoebaNet [68] is applied for high accuracy while light-weight backbone such as MobileNet [69], ShuffleNet [70], or SqueezeNet [71] is applied for time efficiency or deploying on lightweight devices.

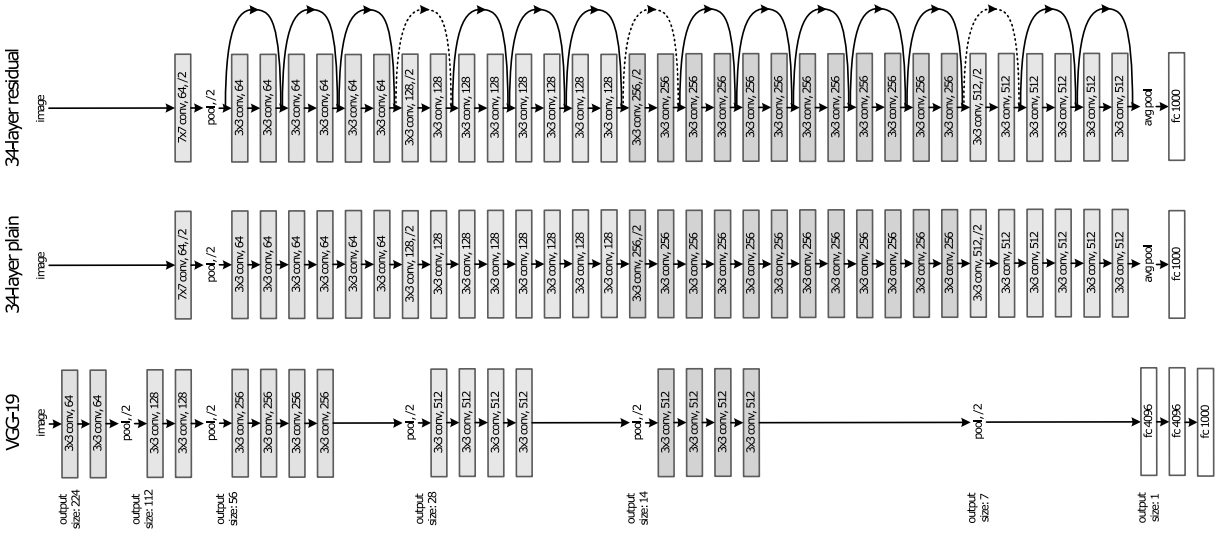


Figure 2.2: Architectures of some backbone networks used in detection networks.

Selective Search [72; 73] methods are applied to produce the spare set of candidate proposal in the first two-stage detectors. Region-based Convolutional Neural Networks (R-CNN) [1] applied a convolutional neural network to the second stage to improve the accuracy of the detector. Faster R-CNN [2] integrates Region Proposal Networks with the second stage into a single convolution network to boost the speed and accuracy. Several improvements have been proposed for R-CNN framework [74; 75; 76; 4; 66].

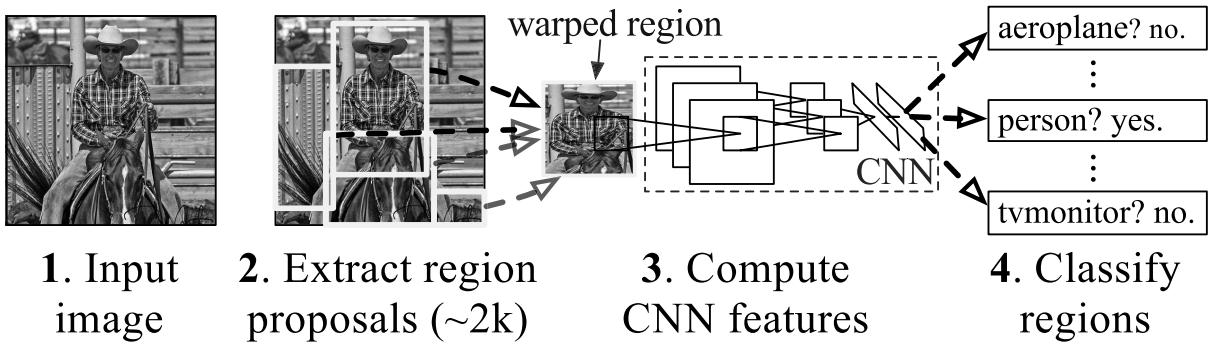


Figure 2.3: Region-based Convolytional Neural Network [1]

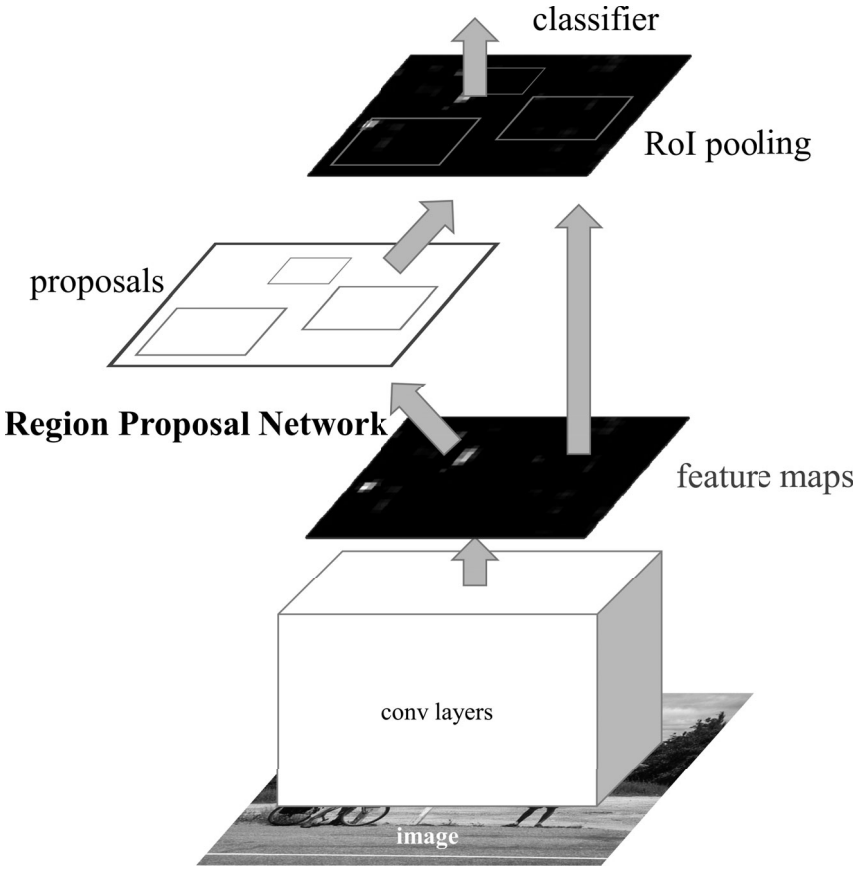


Figure 2.4: Faster Region-based Convolytional Neural Network [2]

Sermanet et al. proposed OverFeat [77] as the first one-stage object detector network. Single Shot Detector (SSD) [78; 79] and YOLO [80; 81] are proposed for the real-time detection. RetinaNet [3], which is widely applied for protozoa detection, derived from Region-based Convolutional Neural Networks (RCNN) [1]. RetinaNet [3] uses a backbone network to extract the feature from the input image and Feature Pyramid Network (FPN) [74] to represent the feature at multiple scale levels. RetinaNet achieves the state-of-the-art in detection taks on COCO benchmark [82] which contains daily life objects and its contexts. However, deep networks haven't succeeded in the domains that have a small number of training samples.

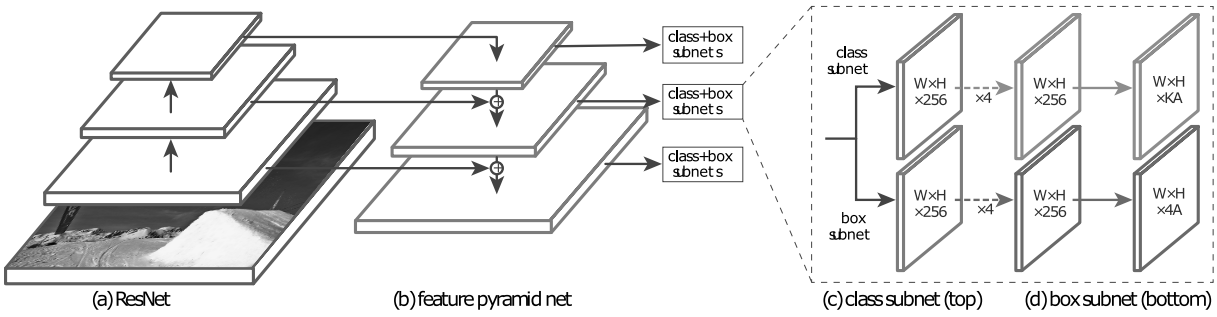


Figure 2.5: RetinaNet [3]

Mask RCNN [4] is also derived from RCNN to perform segmentation at the instance level. Instance level segmentation is to predict whether a pixel belongs to a specific instance or background. Mask RCNN predicts segmentation mask patches corresponding to bounding boxes of detection task. The final segmentation result is the combination of all the patches. Since this study applies segmentation information only to enhance the accuracy of the identification task, the model only performs semantic segmentation which tries to predict whether a pixel belongs to protozoa or background. SharpMask [83] proposed a segmentation refinement technique that takes outputs of all the convolutional layers to produce a sharp segmentation result. The proposed network in this study only refers to the outputs of blocks of the backbone to reduce the number of parameters. Different from Mask RCNN and SharpMask where the segmented masks of the objects take major proportions in the patch, regions of protozoa only take small proportions in the image. In the scenario that the background area is much larger than the areas of

the objects, the network just needs to predict "background" for all the pixels to get high accuracy. Therefore, the proposed network has to deal with the case of predicting all "zero". This study employs ResNet50 [66] (which groups its layers into 4 blocks) as the backbone network for RetinaNet and the proposed model. The proposed network is built up from RetinaNet to inherit the capability of detection. A segmentation network is implemented into RetinaNet to produce a segmentation mask for the input image. The segmentation mask is applied to filter out the background and feed the filtered image to RetinaNet again to identify the species of protozoa in the image. For DNA profiling image dataset, this study introduces Attention-driven RetinaNet that predicts segmentation, refines by attention mask and detects on the attention results.

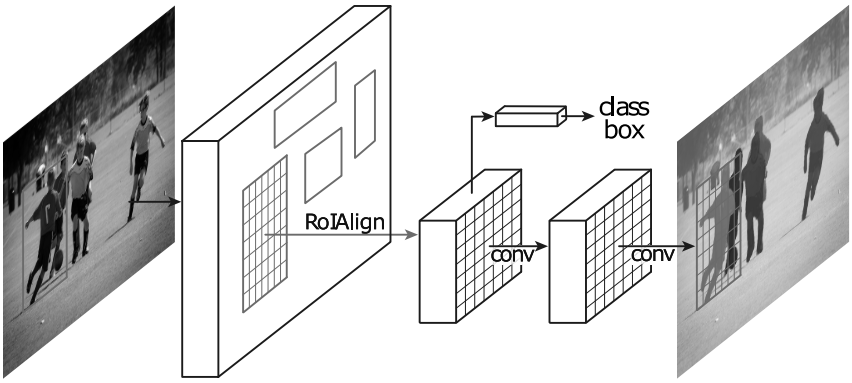


Figure 2.6: Mask RCNN [4]

Chapter 3

Finding the characteristic features

3.1 Characteristic Features for Object Detection

In recent years, a large number of objects can be detected successfully by Deep learning-based methods. In large-scale datasets, Deep Learning-based methods can achieve human-level performances for general daily life objects. Daily life objects generally have a rich source of features. They can be decomposed into plenty of parts. There are parts that only one category has. Detectors can choose one of those parts as a distinctive feature to distinguish from other categories. However, performing object detection on medical image domains is still a challenging problem. The appearances of objects in medical images are much simpler compared to daily life objects. Moreover, Deep Learning-based methods are thrusting for a large amount of training data to achieve good performance. Performing object detection with small training sets is still a remaining challenge. This study assumes that the target objects are irrelevant to the background. Otherwise, background information can be considered a characteristic feature. In the case of a small training set, a particular background or surrounding object may have a high probability of appearing accompanied by the target object. For example, if there are only 5 samples per category, a particular background will have at least $\frac{1}{5} = 20\%$ along with the target objects. However, one specimen may contain multiple species of protozoa. Therefore, the background should be considered as irrelevant to the target objects. To achieve good detection performance,

this study aims to find the characteristic feature of the target objects.

This study takes protozoa as an example for target objects with less visual information. Protozoans are a group of single-celled eukaryotic organisms. Since they can be found in the natural environment, they are clustered into Environmental Microorganisms (EM). They can infect and live parasitically in human or animals bodies. They may cause diseases such as intestinal diseases, diarrhea, muscle pain, transitory edema. They also cause damages to the eyes, brain, or other organs when living parasitically inside the human body. Traditional methods, such as chemical methods or physical spectrum, are used by biologists to detect the presence of protozoans based on their unique responses. For example, Iodamoeba is luminesced when adding iodine. Cyclospora and Sarcocystis are highlighted under UV light. Another efficient way is to apply molecular biology based on DNA or RNA to detect the presence of protozoa. However, the above methods are time-consuming and require expensive equipment. Morphological methods, which identify protozoans by observing them under microscopes manually, are an alternative solution to identify their species. However, distinguishing thousands of species of protozoans is still difficult even for experienced doctors. Automatic methods for detecting, segmenting, and identifying protozoa from micrographs are in demand to reduce the cost and processing time from weeks to a few hours of all the processes. Previous methods on micrographs used shape features to capture morphological differences among species. The regions of protozoa instances are manually segmented out first. The shape features are then extracted such as ellipticity [16; 84], perimeter and area [15; 14], color density information [10], local feature on various color channel [20] or a combination of shape features [10]. The Support Vector Machine [19], Neural Network [18], k Nearest Neighbor [85], Conditional Random Field [22] can be used as the classifier to identify the species.

This study focuses on detecting *Giardia lamblia* (*Gla*), *Iodamoeba butschilii* (*Ibu*), *Toxoplasma gondi* (*Tgo*), *Cyclospora cayetanensis* (*Cca*), *Balantidium coli* (*Bco*), *Sarcocystis* (*Sar*), *Cystoisospora belli* (*Cbe*) and *Acanthamoeba* (*Aca*) (Fig. 3.1). Those species cause diseases to humans but are difficult to distinguish. It is important to identify them to perform the right treatments. Their shapes are similar rounded that make it difficult to

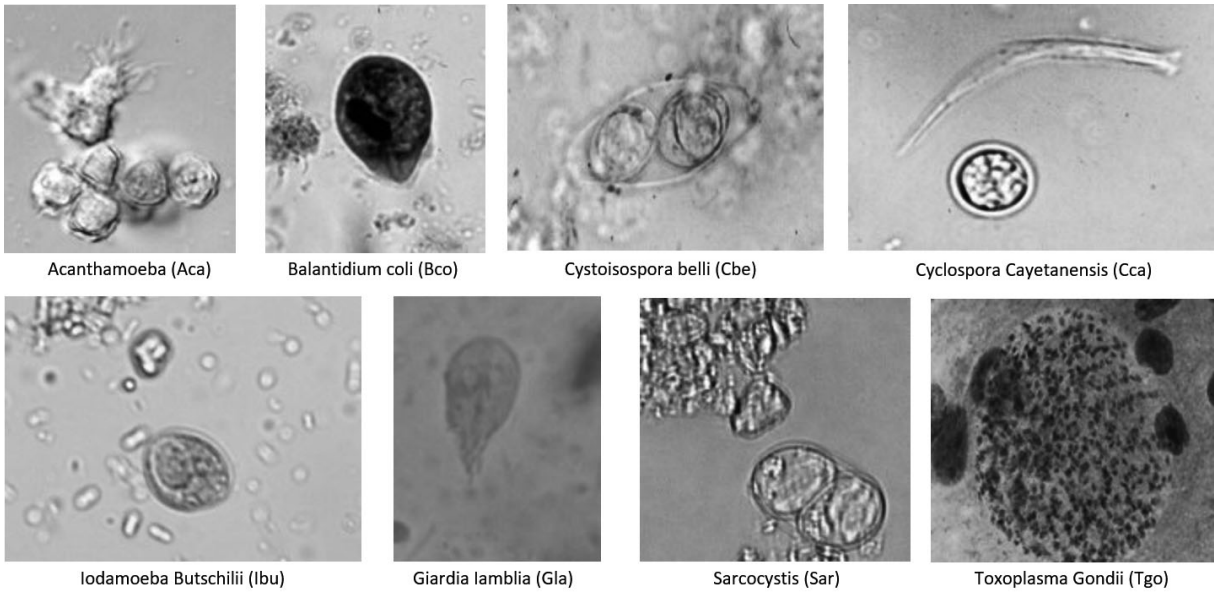


Figure 3.1: Example of the target species.

identify for even experts. The size of protozoa ranges from 1 to 150 micrometers. The range of the size of a species may overlap with that of another. However, the size of protozoa in the micrographs also depend on the zoom scale levels. Therefore, methods that are based on the sizes and the outer shape appearances are inefficient on these species. The protozoa are transparent and crystal-clear with no color. To highlight their appearance, medical institutes apply staining to enhance the contrast in the micrographs. The color information of a certain species in micrographs changes depending on the staining methods (Fig. 3.2). The staining methods used by medical institutions may be different. Therefore, micrographs collected of the same species differ in color conditions. It is necessary to consider methods that are robust to color information to improve versatility. This study aims to find feature representations that are sufficient for protozoa detection, segmentation, and identification and applicable to other species.

This study observes the behaviors of RetinaNet on dataset provided by Dr. Masaharu Tokoro at Kanazawa University to detect protozoa. By applying Class Activation Mapping (CAM) method [86], it can be found that the background information is affects identification predictions (Fig. 3.3). Less than 50% of the area of important regions are the protozoa regions. The background information makes a higher proportion of contri-

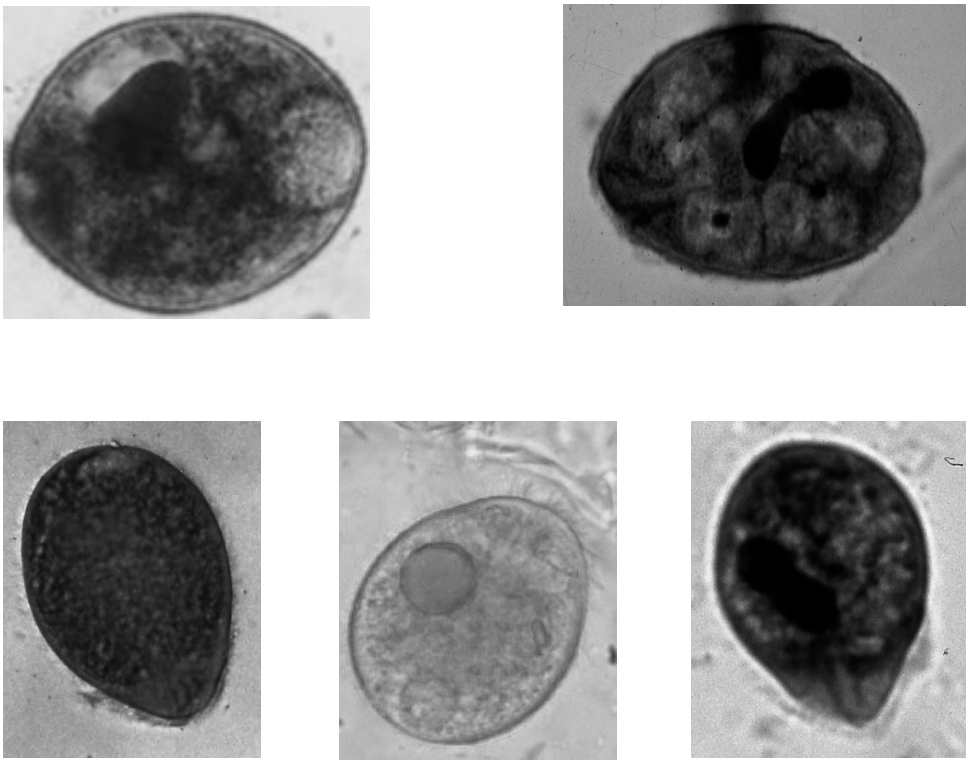


Figure 3.2: An example of the effects of dyeing method on micrographs. *Balantidium coli* instances have different colors in different dyeing method.

bution in making the species predictions. An explanation for this phenomenon is that in the case of lack of data, the network may find that a specific protozoa species often appear with some particular surrounding objects. This study only trains the network with a maximum of 5 samples per species. In the condition of the lack of data, there are often the cases where all the bounding boxes in the training samples contain surrounding objects. A specific surrounding object in the background may occupy at least 20% to appear along with the target species in the dataset. The micrographs of the protozoa are often noisy and contain a lot of dirt, blood cells, stools, etc. In many cases, it is almost impossible to separate the protozoa from those surrounding objects by using rectangular bounding boxes. Moreover, since the datasets are collected from a few medical institutes, background information with a staining method highly correlated with the target species. The insight that a species corresponds to a specific background may be true when the protozoa live in the natural environment. On the other hand, this insight does not hold when they are obtained from the same patient sample. A patient may be infected by

more than one species of protozoa. Multiple species of protozoa may appear in the same micrograph. Therefore, the background information should be excluded when identifying the species.

This study aims to find the characteristic features of the target objects. The network needs to be designed so that the predictions is independent from the background. A solution is to use segmentation to separate the target protozoa instances from surrounding objects.

3.2 Segmentation-driven mechanism

In this study, characteristic features of the target objects are focused on the fly by a Segmentation-driven mechanism. This mechanism guides the deep networks to focus on the characteristic features in the inner texture of the target objects by applying a segmentation mask to eliminate the background of the input image. This section describes how to build the Segmentation-driven mechanism on RetinaNet in detail.

This section describes the segmentation network since the mechanism requires a segmentation mask. Then, the overall architecture of the detection network with the Segmentation-driven mechanism is described. Finally, this study shows the techniques that help to overcome the lack of data problem in the protozoa domain.

The proposed detection network is named Segmentation-driven RetinaNet, which can automatically detect, segment, and identify the protozoa species in the micrographs. The proposed network applied RetinaNet to predict the bounding boxes of protozoa in the micrographs. An instance of RetinaNet without classification layers is used as the backbone to extract the features of the input image. A segmentation network is then built on top of that RetinaNet instance to predict the segmentation masks for the input images. The segmentation network mainly decodes the features extracted by the backbone into single-channel images for segmentation instead of backing to 3 channel images. The segmentation mask is applied to eliminate the background of the input image. Another instance of RetinaNet with classification layers is applied as the detection network to

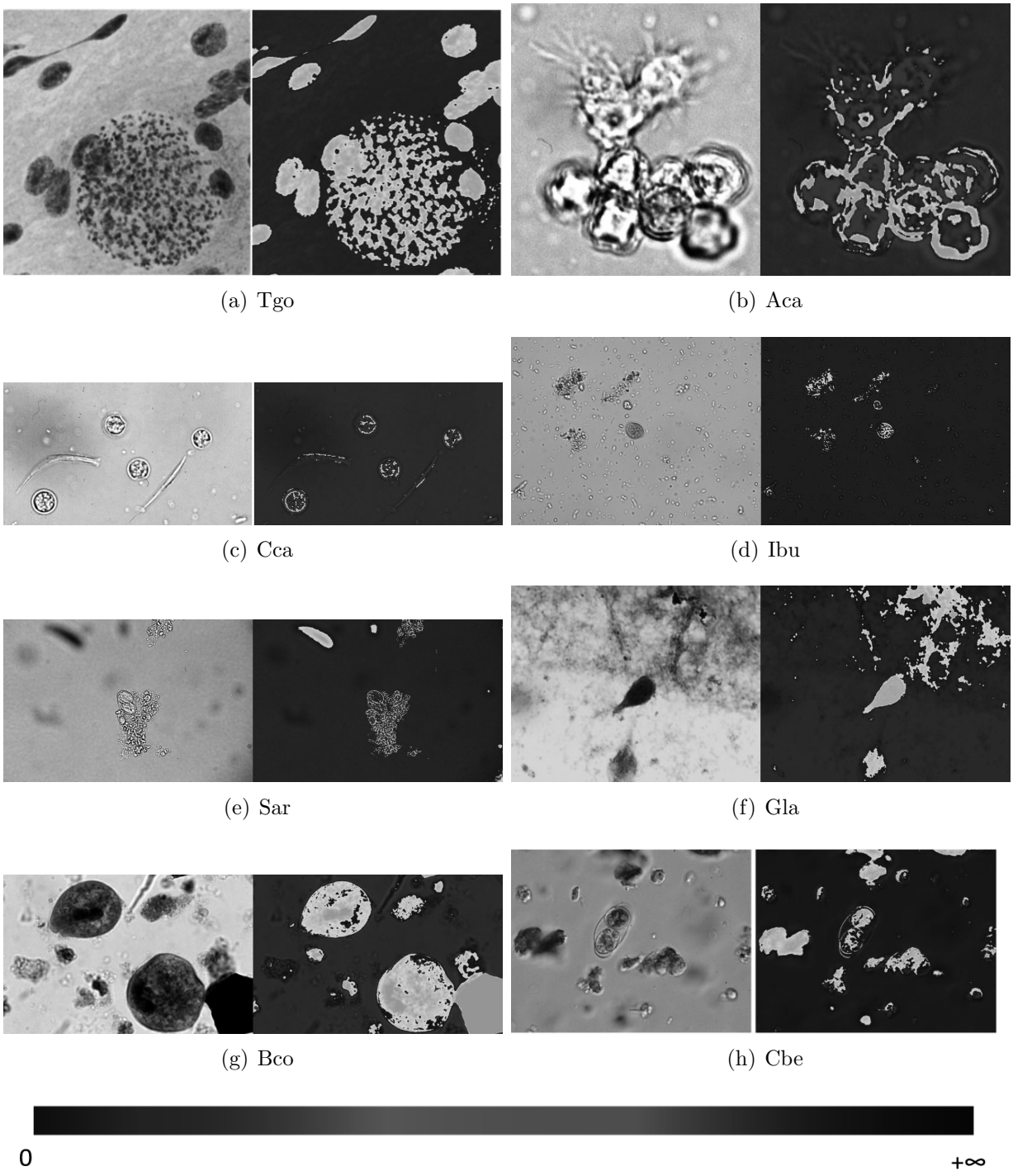
detect the target objects in the filtered image.

A pre-trained weight of RetinaNet is used and fine-tuned to overcome the issue of small number of data. The samples of each species are also divided into multiple groups corresponding to their life cycle stages because there are significant differences among these stages. Dividing images into life cycle stages make it easier to capture the general feature of each life stage and distinctive features from the corresponding stages of other species. The life cycle stage predictions are then mapped into species predictions to obtain the final results.

3.2.1 Segmentation Network

The segmentation-driven mechanism requires segmentation masks to guide the network to focus on the characteristic features of the target objects. A segmentation network is employed to produce the segmentation mask by following the encoder-decoder manner. A backbone network is applied to extract the image features. A Feature Pyramid Network (FPN) is followed to represent the image features in multiple resolutions. The segmentation network is placed on top of the FPN and tries to produce a segmentation mask from features extracted by the backbone network and FPN. To that end, the structure of the segmentation network is a reversion of the backbone network. Convolutional, pooling, and padding layers are replaced with deconvolutional, upsampling, and cropping layers. Figure 3.4 shows the architecture of the segmentaion network. In Figure 3.4, the blocks C1', C2', C3', C4', C5', and FPN' are built as the reversed version of the blocks C1, C2, C3, C4, C5, and FPN, respectively.

Since the extracted feature map from the encoder-decoder manner is compact, the produced segmentation mask is coarse and blurred. To refine the segmentation results, this study follows the idea of a refinement segmentation network of SharpMask that maps the outputs from later to earlier layers. Unlike SharpMask, this study refers to the outputs of the backbone network (C2 to C5 in the ResNet case) to reduce the number of parameters. This architecture helps the segmentation network obtain information from later to earlier



(a) Tgo (b) Aca

(c) Cca (d) Ibu

(e) Sar (f) Gla

(g) Bco (h) Cbe

(i) CAM value

Figure 3.3: Example results of CAM on RetinaNet. The input images are on the left and CAM results are on the right.

blocks to create a sharper segmentation image. This architecture is compatible with various families of Convolutional Neural Networks. To deblur the segmentation result,

residual connections are added from C2, C3, C4, C5 to C2', C3', C4', C5' respectively. The sigmoid activation function is applied at the last layer to scale the predicted values to the $[0, 1]$ range.

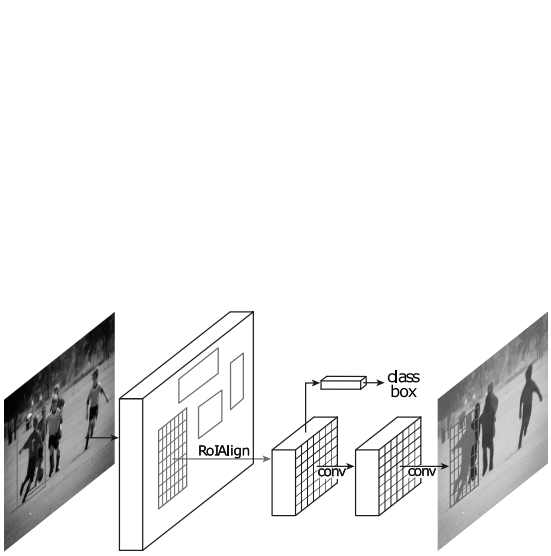
3.2.2 Segmentation-driven RetinaNet

A unified network is built to segment, detect and identify the target objects in the input image. The network predicts the segmentation mask for the input image, then predicts bounding boxes and identifies the categories. The segmentation mask is applied to filter out the backgrounds to force the network to take only target object instances into account when producing the identification predictions.

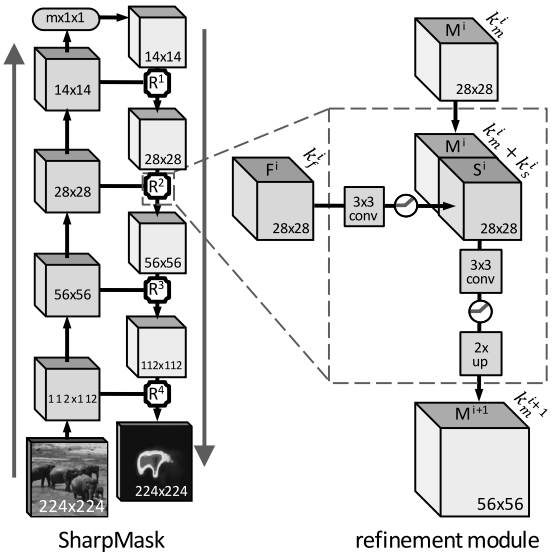
The architecture of the proposed network is described in Fig. 3.5. This architecture consists of 2 instances of the RetinaNet, but they are assigned different weights to deal with various task. The first instance of RetinaNet is employed as the backbone for the segmentation network as described above. The input image is fed into the first RetinaNet instance to predict bounding boxes of the protozoa in the micrograph. The first RetinaNet instance predicts the bounding boxes since it requires background information to indicate the location of protozoa in the environment. The extracted feature by the FPN of the first RetinaNet instance is fed into the segmentation network to produce the predicted segmentation mask. The network computes the element-wise multiply between the input image and the predicted segmentation mask to filter out the background in the input image. Then the filtered image contains only the target objects' instances. It is then fed into the second RetinaNet instance to predict the probabilities of species for the bounding boxes. Since the segmentation mask is applied to guide the network in producing the predictions, the proposed network is named Segmentation-driven RetinaNet.

3.2.3 Data Augmentation

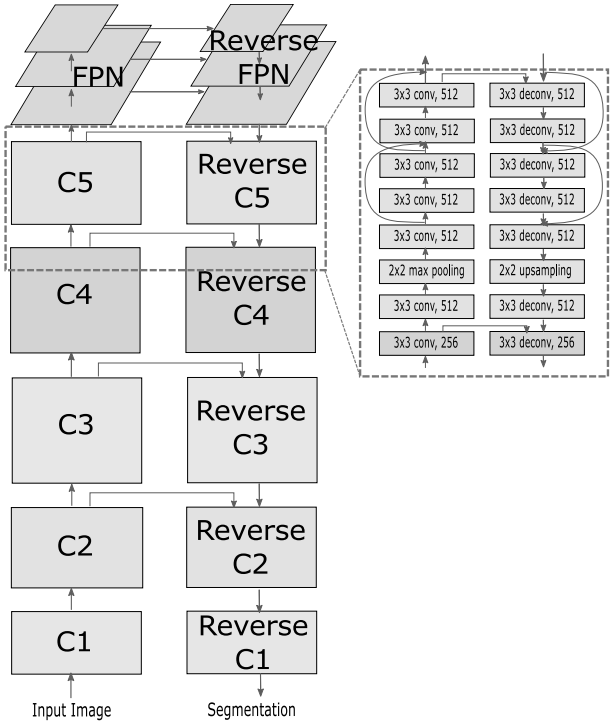
The small number of data is one significant difficulty in the detection task of objects with less visual information, especially in the medical image domains. This section considers



(a) Mask RCNN [4]



(b) Sharp Mask [83]



(c) The Segmentation Network

Figure 3.4: The architecture of the proposed segmentation network with ResNet50 as the backbone.

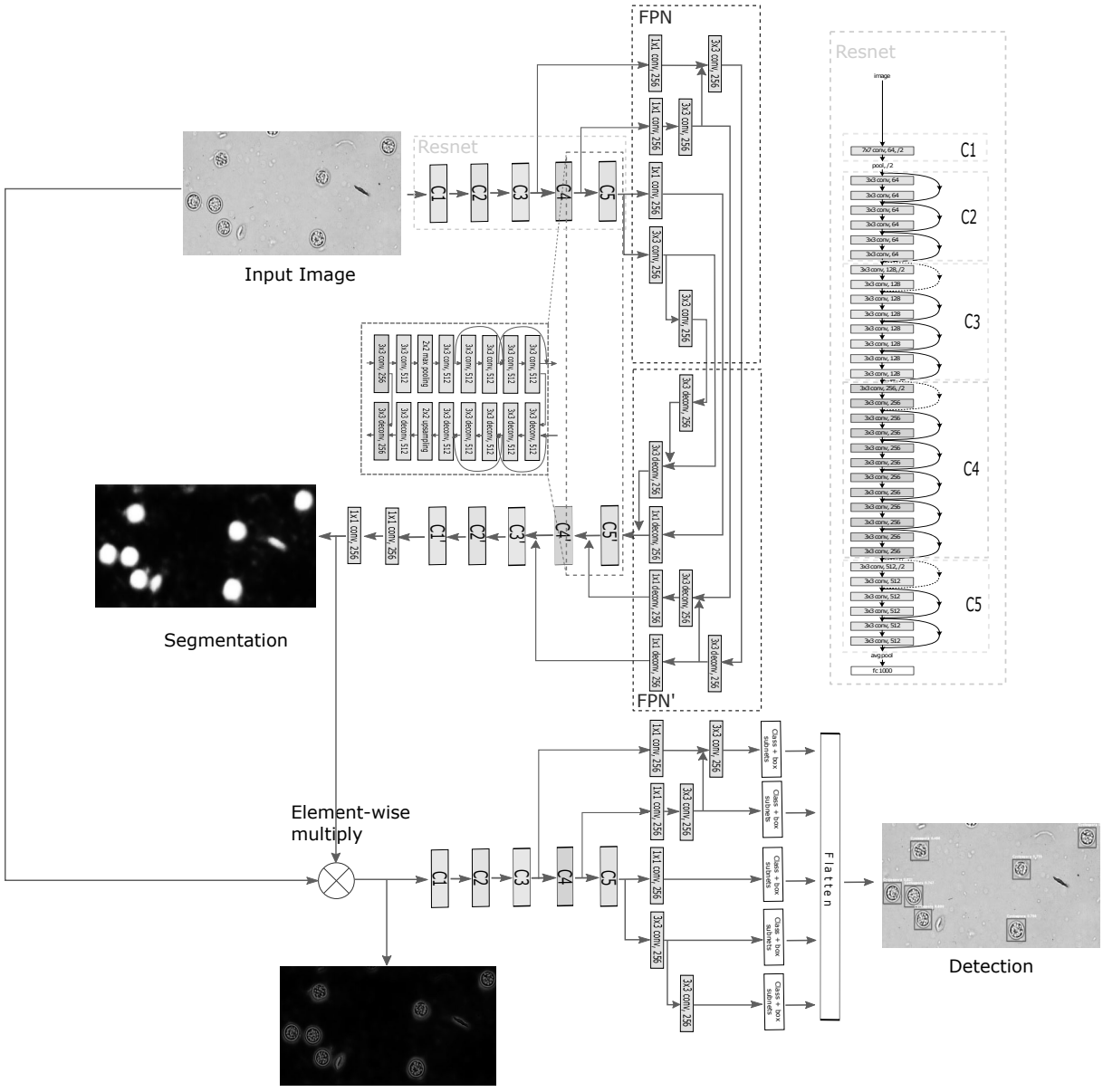


Figure 3.5: The proposed Segmentation-driven RetinaNet.

the protozoa domain as an example of objects with less visual information. It takes time and effort to obtain the samples from the environment. Since the target species cause human diseases, there are new samples only when there is a new patient. Several data augmentation techniques are applied to enhance the data to overcome the small number of data problems.

Rotation: In the micrographs, the protozoa instances appear in arbitrary orientations. Based on this property, the protozoa regions can be rotated to get more data for training without creating unrealistic samples. The centers of the bounding boxes of protozoa instances are chosen alternatively as the origin of coordinates. The image is then rotated around that center in all directions with an interval of 10 degrees. The new bounding boxes of protozoa instances are calculated concerning the rotation angle in the new rotated images.

Color Transfer: The color distributions of protozoa instances in the micrographs may appear completely different due to various staining methods. The data is augmented by changing the images' overall color distributions to enhance the model's robustness on color changes. The color transfer technique [87] is applied to simulate various staining conditions. To stimulate to the target staining condition, its color distribution is required. Color distribution images, which contain only the background of the micrograph of the target staining conditions, is prepared to calculate the target distribution. The protozoa instances on the chosen images are removed to obtain the color distributions of the targeted staining methods. In this study, 21 target color distribution images are created for color transfer regarding the micrographs' different staining and lighting conditions (Fig. 3.6).

3.2.4 Segmentation Loss

The total loss for training the entire model is as follows:

$$L = L_{cls} + L_{box} + L_{seg}, \tag{3.1}$$

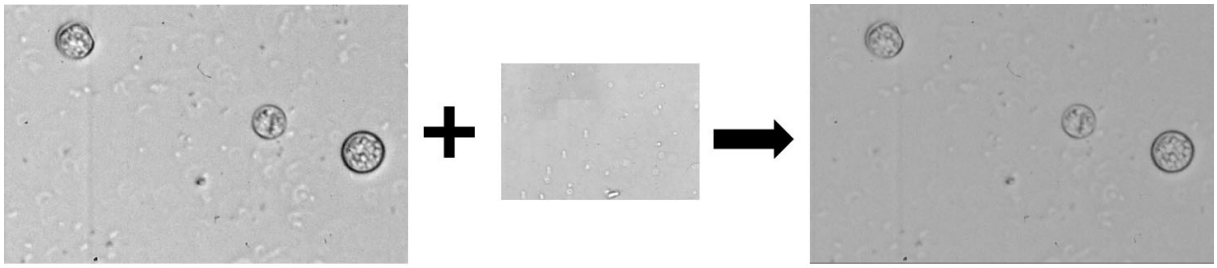


Figure 3.6: An example of color transfer technique to simulate the staining method on the micrographs of *Cyclospora cayetanensis*.

where L_{cls} is the identification loss, L_{box} is the bounding box prediction loss, and L_{seg} is the segmentation loss. This study applies the focal loss for L_{cls} and the smoothed absolute value loss for L_{box} that are identical to RetinaNet.

The weighted cross-entropy loss function is applied to train the network for the segmentation task. In the micrographs, the protozoa instances only take a small proportion. The majority of the segmentation ground truth is background. Non-weighted loss functions often result in the trivial case that the model only predicts "zero" for all the pixels. To balance the contributions of foreground and background pixels, the weight for foreground pixels should be applied. This research applies the weight for background pixel as $w_b = 1$ and the weight for foreground pixel as

$$w_f = \frac{\#background\_pixel}{\#foreground\_pixel} \tag{3.2}$$

which corresponds to the ratio between the number of foreground pixels and that of background pixels. The segmentation loss on one image is as follows:

$$L_{seg} = -w_f \sum_{i,j,s_{i,j}=1}^{w,h} log(\hat{s}_{i,j}) - w_b \sum_{i,j,s_{i,j}=0}^{w,h} log(1 - \hat{s}_{i,j}), \tag{3.3}$$

where w and h are the width and height of the image, s is the segmentation ground truth, and $\hat{s}$ is the segmentation prediction, respectively. This segmentation loss only focuses on the binary class segmentation problem where 0 and 1 correspond to background and

foreground, respectively, instead of multiple categories or instances.

3.2.5 Network Training

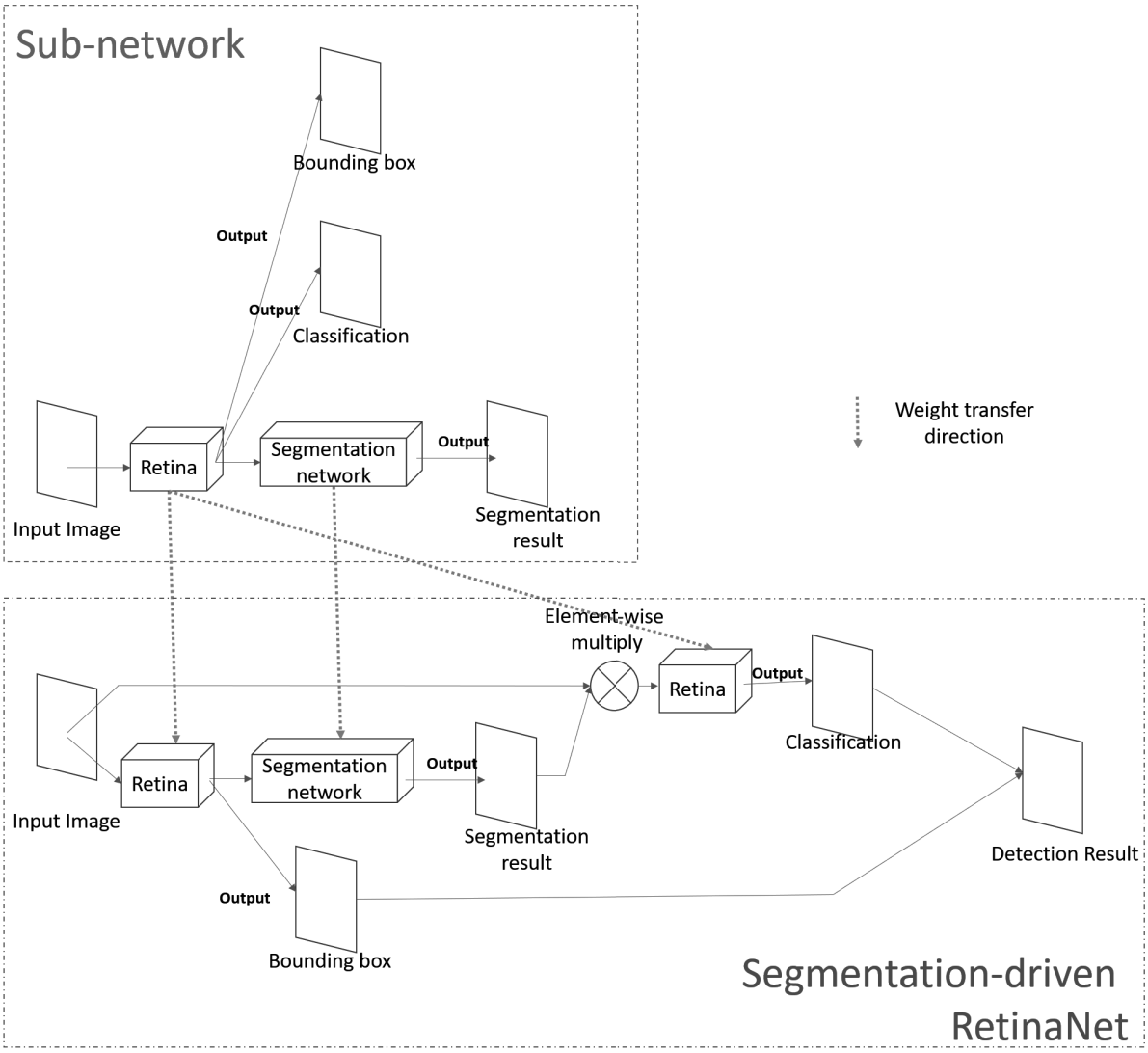


Figure 3.7: Network training procedure.

To train the Segmentation-driven RetinaNet for the protozoa domain, a pre-trained weight of RetinaNet is applied with the resnet50 as backbone trained on COCO dataset [82]. The transfer learning technique is affective for cases where only small number of dataset available (Fig. 3.7).

A sub-network that combines the first RetinaNet instance and the proposed segmentation network is created. In this sub-network, the RetinaNet instance also predicts the

categories of objects inside the bounding boxes. The weight of the pre-trained model is transferred to the RetinaNet instance. Since the segmentation network is not pre-trained, its derivatives concerning the weights are higher than the pre-trained Retina's. A small loss-weight for segmentation loss is applied to prevent the network from excessively modifying weights that could depress detection and identification performance.

The proposed Segmentation-driven RetinaNet is trained with the weight of the above sub-network. The weight of the segmentation network is transferred from the trained sub-network. The weights of both RetinaNet instances are initialized to the weights of RetinaNet in the sub-network. The three tasks are trained one by one alternatively. While training one task, the learning rates are set for the others two to be small instead of zero, so the network learns the weights adapting to the three tasks.

3.3 Evaluation

3.3.1 Dataset

Experiments are conducted on the protozoa dataset provided by Dr. Masaharu Tokoro at Kanazawa University, consisting of 38 images with 43 samples for training and 31 images with 74 samples for testing. The image resolution was not constant, but ranging from 824x941 to 4086x1725 pix. All the images are real cases taken from the patients. There are one to multiple instances in an image. The numbers of instances for each life-cycle stage in training data are given in Table 3.1.

3.3.2 Evaluation Metric

Since the proposed network only performs segmentation at the class level, binary accuracy, precision, and recall metrics are applied to evaluate the segmentation performance of the proposed models. To evaluate the performance of the proposed network on detection and identification, mAP metric is from PASCAL VOC [88] is applied. The mAP is the interpolated average precision designed to penalize for missing target object instances,

Table 3.1: Number of training samples for each life-cycle stage.

	$\#trainingsamples$
Cca oocyst 1	3
Cca oocyst 2	2
Cca oocyst 3	1
Sar oocyst 1	1
Sar oocyst 2	2
Aca	5
Ibu	3
Tgo	4
Gla	3
Bco cyst	2
Bco trophozoite 1	3
Bco trophozoite 2	4
Cbe oocyst 1	3
Cbe oocyst 2	3
Cbe oocyst 3	2
Cbe oocyst 4	2
Total	43

duplicated detections, and false-positive detections. To calculate the mAP value, the detection outputs of the model are sorted by the descendant order of confidence. The accumulated precision and recall are computed following the descendant order. Finally, the average precision is computed corresponding to the given set of recalls. In PASCAL VOC, the set of chosen recall is the range [0,1] with the interval of 0.1. The formula for mAP of PASCAL VOC is given as follows:

$$mAP = \frac{1}{11}\sum_{t=0}^{10} max\{precision_i | recall_i >= t\}. \tag{3.4}$$

where $precision_i$ and $recall_i$ is computed at the i^th detection.

3.3.3 Result of Experiment

All the experiments are performed on a Xeon Gold 6130 2.1GHz CPU, 128 GB RAM, and 2 NVIDIA Tesla P100 GPUs. The version of the proposed model was implemented

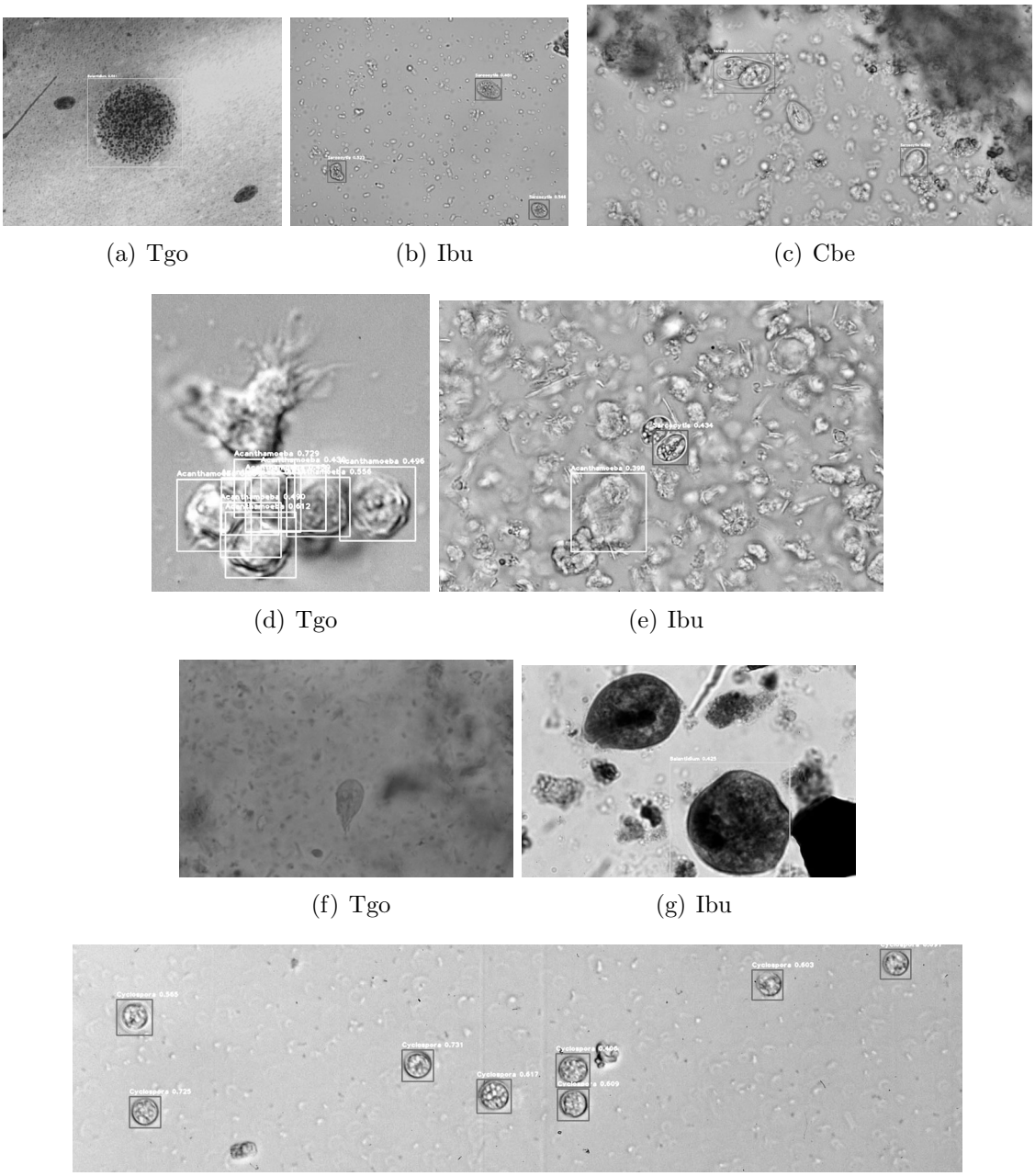
Table 3.2: mAP, precision, and recall with respect to species.

	Cca	Sar	Aca	Ibu	Tgo	Gla	Bco	Cbe	avg.
mAP									
RetinaNet	0.82	0.10	0.53	0.18	0.00	0.36	0.39	0.00	0.30
Sub-network	0.86	0.18	0.78	0.73	1.00	1.00	1.00	0.32	0.73
Proposed method	0.89	0.50	0.73	0.84	1.00	0.84	0.61	0.75	0.77
Precision									
RetinaNet	1.0	0.14	0.48	1.00	0.00	0.50	0.50	0.00	0.45
Sub-network	0.94	0.25	0.74	1.00	1.00	1.00	1.00	0.50	0.80
Proposed method	0.97	0.50	0.71	0.90	1.0	0.75	0.83	0.60	0.78
Recall									
RetinaNet	0.83	0.67	0.94	0.10	0.00	0.33	0.43	0.00	0.41
Sub-network	0.94	0.67	1.00	0.70	1.00	1.00	1.00	0.67	0.87
Proposed method	0.92	0.67	1.00	0.90	1.00	1.00	0.71	1.00	0.90

using RetinaNet. The score threshold for making the decision is 0.34. The batch size is set as 2 for training. The overlapping threshold for a bounding box to be matched to a ground truth box is 0.8.

The colors of the predicted bounding box are pink, red, white, blue, yellow, black, light blue, and green for *Cca*, *Sar*, *Aca*, *Ibu*, *Tgo*, *Gla*, *Bco*, and *Cbe*, respectively. To visualize the segmentation results, the green channel of input images is replaced with the predicted segmentation masks.

One mistake may cause a significant decrease in the result due to the lack of data problem. The original RetinaNet and the proposed networks capture all the protozoa instances. However, there are misidentifications of the species, especially between *Ibu* and *Sar*. One misidentified instance will negatively affect the precision of the predicted category and the recall of the true category. In these three examples (d, e, f) in Fig. 3.8, RetinaNet successfully localized the protozoa instances but failed to identify the species. The false identifications hurt the precision of the predicted categories (*Ibu* and *Bgo*) and the recall of the true categories (*Tgo*, *Ibu*, and *Cbe*).



(a) Tgo (b) Ibu (c) Cbe

(d) Tgo (e) Ibu

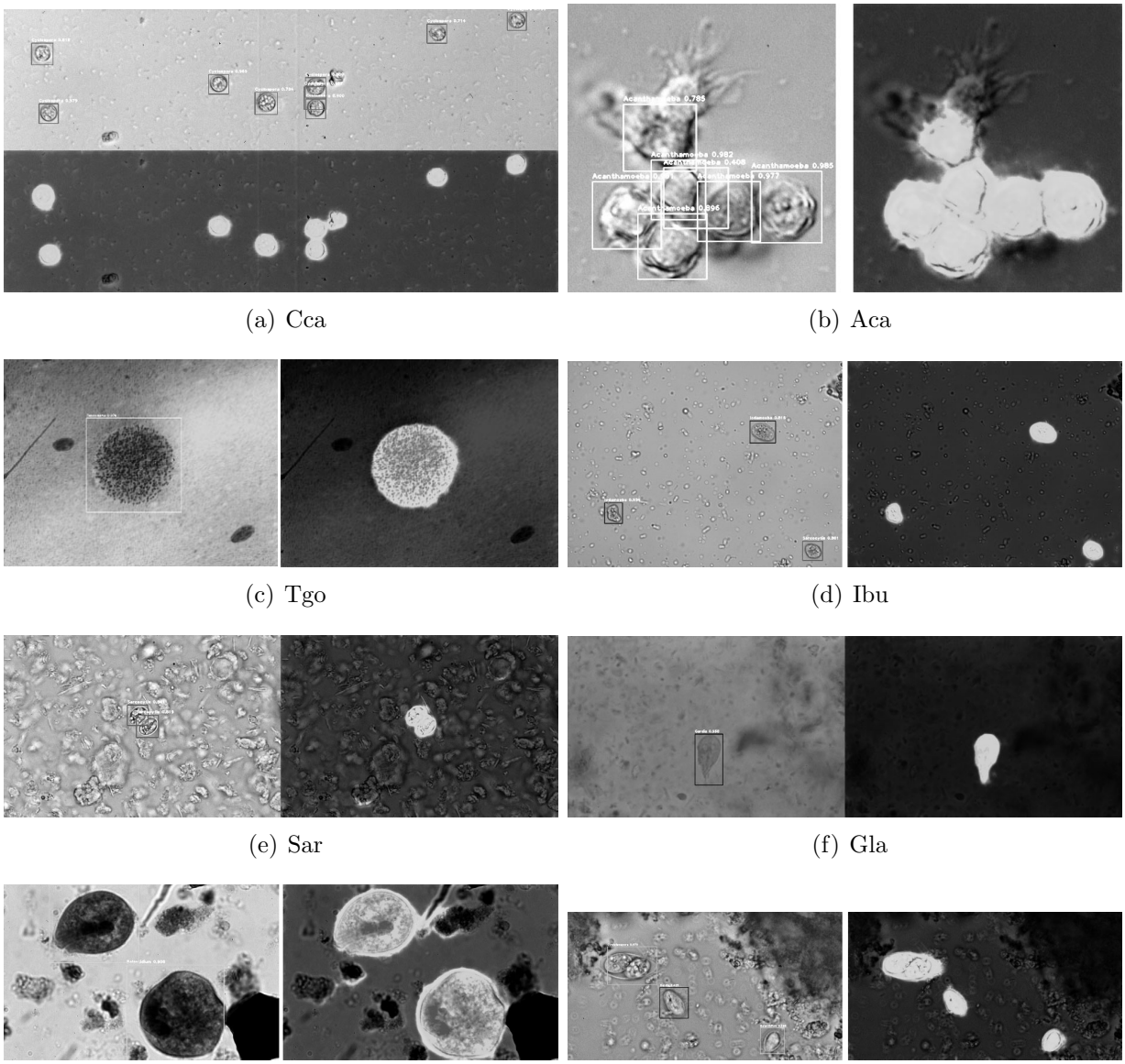
(f) Tgo (g) Ibu

(h) Cbe

Figure 3.8: Examples of detection of the original RetinaNet.

3.4 Summary of Finding characteristic features

This chapter focuses on finding the characteristic features for objects with less visual information in small datasets. In this problem, it can be found that the deep learning-based Object Detection methods utilize the background information to make the identification prediction. It is insufficient when the objects appear in other environments or objects



(a) Cca (b) Aca

(c) Tgo (d) Ibu

(e) Sar (f) Gla

(g) Bco (h) Cbe

Figure 3.9: Examples of detection and segmentation of Segmentation-driven RetinaNet.

of different categories appear in the same images. Therefore, this study aims to find the characteristic features of the textures of the target objects. This study introduces the Segmentation-driven mechanism to guide the deep network to find the characteristic features of the target objects. Segmentation masks are applied to filter out the background of the images. By removing background information, only the features in the regions of target object instances are learned.

This idea is demonstrated on the protozoa domain. Segmentation-driven RetinaNet

is proposed to guide the network to take only the target object instance regions in the image into account to find characteristic features for identifying the protozoa. Data augmentation technique is applied to overcome the small number of data problem. This study successfully trains the network to detect, segment, and identify protozoa with high accuracy even though there are at most 5 samples per life-cycle stage for training. Experimental results on the protozoa dataset show the effectiveness of the proposed model over the original RetinaNet.

Chapter 4

Finding the distinctive features

4.1 Distinctive Features for Object Identification

An object can be decomposed into its outer shape, appearance, and inner textures. Those are features that characterize the target objects. The characteristic features help to detect the target objects. The objects in the target object set may share some characteristic features in common. The system aims to learn the unique features of each category to identify. Those features are called distinctive features. In the case of lack of training data, a common feature, in theory, may be considered a unique feature. It is the bottleneck of generalizing tasks for new input samples. Therefore, choosing quality distinctive features is essential to detecting and identifying new test input.

Color images have three color channels (red, green, and blue). In each channel, a pixel takes a value from 0 to 255. With three color channels, a pixel of a color image has 255^3 possible values. The organizations of the inner pixels define the textures of the target objects. The texture space of the objects in color images is enormous. On the other hand, domains captured by specific devices are grayscale images with one color channel. With a single color channel, a pixel of a grayscale image only has 256 different values. Detecting objects with less visual information in grayscale is a challenging problem. The visual information in grayscale images is much less than the color images. Objects in grayscale images are characterized by their outer shape, pixel intensities, and connectivity. The

outer shapes of the objects contribute more to the identification than the texture features.

This chapter discusses Genome Profiling images as examples of objects with less visual information in grayscale images. The attention-driven mechanism is applied in place of the Segmentation-driven to find quality distinctive features for the identification tasks of both the protozoa and Genome Profiling image domains. Protozoa domain and Genome Profiling domain may require different prerequisite preprocessing steps. For the Genome Profiling images, the attention mechanism enhances the outer shape of the target trajectories and spiddos. On the other hand, the attention mechanism enhances the selective mask for the target objects' instances for the protozoa domain.

4.2 Prerequisite Preprocessing Step

4.2.1 Genome Profiling domain

This subsection aims to reduce noise in the images and emphasize the thin trajectories whose intensities are low. The intensities of pixels of a trajectory are not constant. Within a trajectory, the intensities are lower around the bottom and higher around the top of the image. The thinner trajectories have lower intensities than the thicker trajectories. Global normalization, which change all the pixels into the same range of intensity values, may remove the thin trajectories. Therefore, the input image is divided into multiple patches to be treated locally. Local equalization is applied to emphasize the trajectories.

Normalizing the trajectories is a challenging problem because their pixel intensities are non-identical because of TGGE nature. Surrounding noise may have higher intensities than lower regions of trajectories prevent globally enhancing methods. An image processing procedure is proposed to locally enhance the intensities of the target trajectories in TGGE images. Morphological operators are performed to erose noise. An iterative updating process is proposed to eliminate isolated regions. At each iteration, the process tries to make a greater difference between the intensity values of the trajectories and the isolated regions. After a certain number of updating iterations, the intensities of the tra-

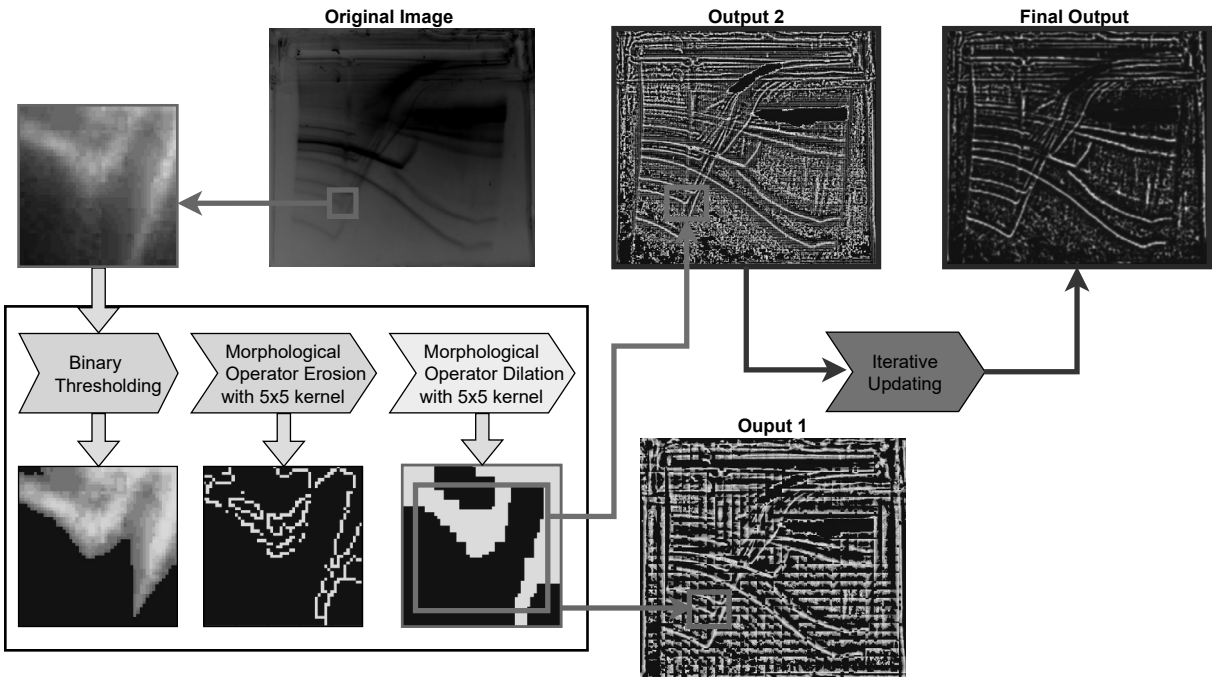


Figure 4.1: Preprocessing procedure.

jectories will be more different from the isolated regions that a threshold can be applied to eliminate the noise.

In a local patch, the pixels of the trajectories have higher intensities than the surrounding pixels. The binary thresholding is applied to remove surrounding noise pixels. However, the thresholding alone cannot separate one trajectory from another that are close to each other. Therefore, morphological operators are applied to remove lower intensity regions between two trajectories. Erosion operator is applied to erode thin trajectories and small regions. This operator is also helpful for separating one trajectory from another that are closed to each others. Dilation operator is applied to fill in the gaps which may be created by erosion operator. However, morphological operators may cause a side effect at the edges of the resulting patches. Dilation operator also fills in the region between trajectories and edges of the patch. It creates vertical and horizontal stripes in the result image when combining all the resulting patches. To avoid this side effect, the edges of the resulting patch is removed.

The isolated small regions, which can be considered as noise dots in the images, need to be removed. To that end, an iterative updating process is applied to emphasize the

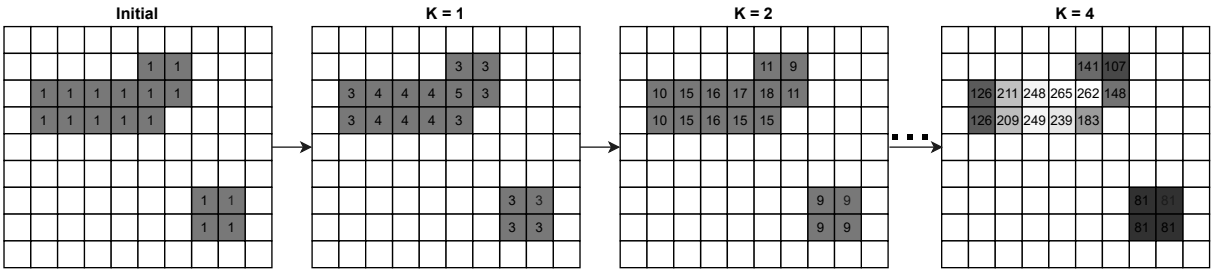


Figure 4.2: An example of the proposed iterative updating method on a line and an isolated dot.

pixels of the lines over the pixels of isolated regions. At the initial stage, pixels are set to 1 if its intensity value is higher than a threshold. Others are set to 0. The values are then updated iteratively. In each iteration, only the pixels, which are initialized as 1, are updated. The value of the pixel is updated as the sum of its value and its neighbors in a kernel of 3x3. The values of pixels that have more connections increase faster than the pixels that have fewer connections. In the next iterations, pixels in large regions spread their high values therefore their values increase faster than pixels in isolated small regions. After a certain number of iterations, the values of pixels in large regions are much higher than pixels in isolated regions. A threshold is applied to remove the isolated regions.

Figure 4.1 illustrates the preprocessing procedure with an example. The input image is divided into multiple image patches. An image patch of two trajectories that are close to each other is shown in the example. Since the pixels surrounding the trajectories have relative high intensities, high threshold value may remove the trajectory pixels. Erosion operator removes the pixels surrounding the trajectories and remains thin lines indicating the positions of the trajectories. In this study, the kernel of 5x5 is applied for the morphological operators. Larger kernels may remove even the target trajectories while smaller kernels remains more surrounding pixels. The dilation operator is then applied to recover the trajectories. $Output1$ in Figure 4.2 is the result of sticking back all the image patches after the morphological operators. Since the dilation operator also fills the gaps near the edges of the image patch, the $Output1$ contains vertical and horizontal strides. To remove the strides, smaller region (pink rectangle) is remained in the resulting patch. $Output2$ is the result when sticking the smaller region of the resulting patches. Therefore,

to get the same resulting patch size from $Output1$, the image patches divided from the input image must have larger size for the $Output2$. The divided patches may overlap each other on the input image. The $FinalOutput$ is obtained after removing the isolated dots in the $Output2$ by applying the iterative updating process.

Figure 4.2 demonstrates the iterative update process with an example that contains a line and a isolated dot. The pixel intensities are set to 1 for both the line and the isolated dot at the initial stage. After the first iteration, the pixels at the center of the line have higher values than the isolated dot. Since the intensities of the line's pixels are higher than isolated dot's, the gap between them becomes larger after every iteration. After several iterations, a threshold can be applied to remove the isolated dot. In the example in Figure 4.2, threshold can be set as 100 to remove the isolated dot.

Object detection methods require manual annotation of the regions of the target objects to separate them from the background in the images. Rectangle bounding boxes are commonly used to indicate the regions of the objects since it is easy for non-experts to annotate. In a small region within a bounding box, the spiddos share similar appearances with uninterested flexion points or noise in the images. Keeping track of the target trajectories is required to distinguish the spiddos from the noise in the images. Since some regions in the top half of the images are in over brightness that occludes the target trajectories, the trajectories are disconnected.

Segmentation can be applied to keep track of the trajectories to overcome the fragmentation issue. The pixels from the beginning to the end of the trajectories can be indicated in the segmentation masks. The segmentation areas can be beyond the bounding boxes of the target objects. Segmentation is applied to keep track of the trajectories even in the occluded regions.

Segmentation for Genome Profiling Images

The goal is to detect the spiddos in the TGGE images. The spiddos are annotated as points in the input images by the domain experts. For general object detection methods, annotations of the target objects are represented as its rectangle bounding boxes to help

the networks to understand what the target objects should look like. To train the detector on the spiddos, the positions of the spiddos are re-annotated as rectangle bounding boxes whose centers are those spiddos coordinates. This helps to encode more information of the trajectories from the left and right of the spiddos than the pixel intensity value at the spiddos coordinate.

Rectangle bounding boxes may contain surrounding isolated regions. Those regions may be considered as noise in the training process. Carefully indicating which pixels are the target trajectories would clear the ambiguity in training samples. Therefore, a segmentation mask, which indicates only the pixels of the target trajectories, is applied to filter out those surrounding regions. Training the network to segment forces the network to learn the ability to remove the noise and focus on the important regions.

Segmentation also helps to keep track of the target trajectories. Due to the over brightness regions in the input images, the target trajectories are disconnected. The segmentation samples for training, which are prepared manually, contain the annotation of the bright pixels that belong to the target trajectories and dark pixels that can connect the disconnected regions of the trajectories. Indicating the dark pixels helps the network learn to predict the trajectories in the occluded regions.

4.2.2 Protozoa domain

In this chapter, protozoa dataset from ICIP 2022 Grand Challenge is being used to demonstrate the idea of Attention-driven mechanism. In Chapter 3, the detection accuracy is improved by applying the segmentation mask to filter out the background information. Manually annotating segmentation for the entire dataset takes a long time and a great effort. A solution for the datasets with no segmentation ground truth automatically generates segmentation masks. It can be done either by:

- **Training a segmentation model in a smaller dataset**. A simple segmentation network can be applied to segment the entire dataset. To reduce the human effort, only a few samples are required to be annotated. Since the microorganisms can

swim freely in their environment, their appearances can be at arbitrary orientations. Therefore, rotating the samples can be applied as data augmentation to train the segmentation network.

- **Applying a segmentation network trained on another similar dataset which has segmentation ground truth**. Since they have circular round shapes, it is unnecessary to have the same target category as the target dataset. However, instance-level segmentation may not work due to the differences in target categories. On the other hand, semantic-level segmentation, which tries to indicate the foreground and background pixels, is suitable for this problem.

This study follows the second solution by applying the trained network described in Chapter 3. Segmentation-driven RetinaNet from Chapter 3 generates semantic segmentation masks that indicate whether a pixel belongs to objects or background. However, the generated segmentation masks may be insufficient since the model is trained on different species or a small number of training samples. Training the model with such roughly generated segmentation masks as ground truth may fail. The insufficient segmentation masks are refined via the attention mechanism in the next stage.

4.3 Attention-driven mechanism

4.3.1 Visual Content Attention Block

The segmentation results are coarse and blurred around the boundaries of the objects. The quality of the segmentation is not efficient enough for the deep networks to extract the shape features. Moreover, the bounding boxes of the spiddos cannot cover entire trajectories. Therefore, it requires global features to map the segmentation of the trajectories to its spiddos. This study proposes to integrate the attention mechanism into the detection network to enhance the trajectories' information via the spiddos regions.

In recent years, the strength of attention is proved in many approaches of Computer Vision, especially Object Detection [89; 90]. The goal of most attention mechanisms is

to intensify the relationship of entities. The influence of attention is no more doubt in most neural approaches. With the support of attention, the hidden and important characteristics of images and objects are emphasized in another viewpoint, which is efficient to deploy in most Deep Learning models. Among many kinds of attention, Multi-head Attention is one of the most potential components in previous approaches. Different from the other ones, Multi-head Attention considers entities in many parallel viewpoints, which is completely useful to put it into practice [91]. Recently, this kind of attention is a rapidly growing tendency in Computer Vision [92]. It is considered to be the dominant Convolution Neural Network in most vision applications. Although it is still too early to know the long-term evolution, there is no denying that Multi-head Attention is effective in many vision approaches.

In this problems, the challenges come from the objects' boundaries. The specific characteristic of the genome is a tough struggle for Convolution Neural Network to digest and emphasize the difference in images. Visual features are compressed and filtered by kernels in many layers. Although this mechanism is useful to highlight the regional features of images, it accidentally accentuates the global information of images. However, in this typical domain, the interaction between segmentation and original images is vitally important to provide more context for optimizing the learning process. Therefore, this study proposes visual content attention block to incorporate the attention information into the convolutional features from the previous modules. With the aid of Multi-head Attention, the proposed mechanism allows segmentation to observe and exploit the visual content in many simultaneous viewpoints. This strength of the proposed approach is clarified in the visualization of Fig. 4.3. Through the attentive features of the Visual Content Attention, the peculiarity of the genome gradually becomes a real standout.

In particular, the proposed block is a combination of three sub-modules as follows:

- Guided-attention via visual content: As mentioned above, visual content is essential to reveal the relationship between the target objects and global information from the segmentation process. Therefore, the strength of Multi-head Attention is

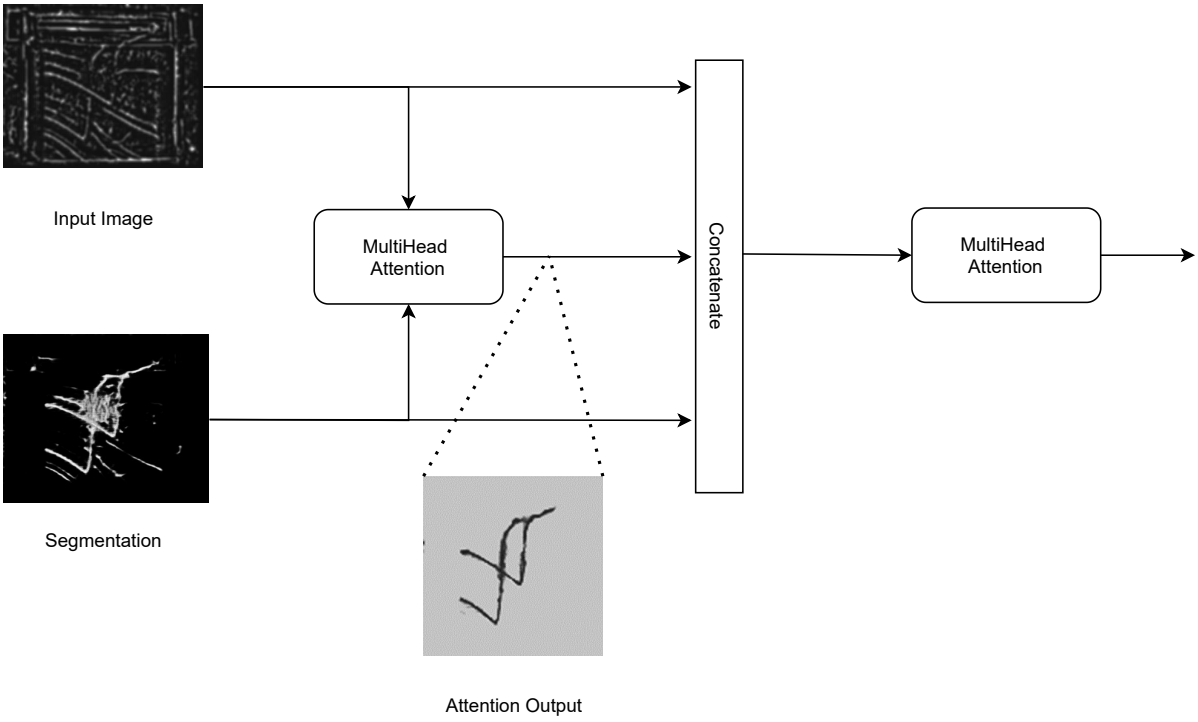


Figure 4.3: Visual Content Attention Block.

applied to gain the attentive scores between images and segmentation of genome. This mechanism allows the proposed model to digest and accentuate the interaction among global and convolution features in many previous layers.

- Residual connection: After enhancing the uniqueness of the genome in segmentation, the previous features are also retained to guide the next self-attention module in feature extraction. Via the residual connection, the global and attentive features of objects are highly preserved in the transmitting process.

- Self-Attention: Finally, the completed signal is intensified by a self-attention module. With the power of self-attention, each feature in the final representation is digested and estimated to reveal its internal importance and relationship against the prediction.

As mentioned above, the module takes advantage of Multi-head Attention in both Guided-Attention and Self-Attention whose difference is based on the configuration of input. Mathematically, with the general inputs, the refined signal is calculated by Equa-

tion 4.1.

$$MultiheadAttention(q, K, V) = [head_1, head_2, ..., head_h]W^O \tag{4.1}$$

Each head $head_j, j \in \{1, ..., h\}$ in Multi-head Attention is pre-defined by users and reflects the various consideration. The detail of each head is pointed out in Equation 4.2. In each specific viewpoint, each query (q) is accumulative by the attentive score from the key (K) and latent information from value V. In most approaches utilizing this attention, the value of K and V is often similar. It allows query features to be considered and intensified in the simultaneously meaningful space.

$$head_j = softmax(\frac{qW_j^Q(KW_j^K)^T}{\sqrt{d}})VW_j^V \tag{4.2}$$

where W_j^Q, W_j^K, W_j^V are the weights for query, key, and value at the $head_j$, respectively.

In the proposed approach, this mechanism makes use of Multi-head attention to accentuate both segmentation and refined features. In the first sub-module, the segmentation information s is updated by its interaction against original images I in Equation 4.3.

$$s^{'} = MultiheadAttention(s, I, I) \tag{4.3}$$

In the next process, the visual content and segmentation are maintained by Residual Transmission via Concatenation operators. Finally, each feature f_i of $f = (I||s^{'}||s)$ from the combination of original images, segmentation, and attentive refinement is escalated by a self-attention module in Equation 4.4.

$$f_i = SelfAttention(f_i) = MultiheadAttention(f_i, f_i, f_i) \tag{4.4}$$

Through the Residual Transmission and Self-Attention in the proposed architecture, it is promising to understand and select the necessary features to enhance the learning and predicting process. With the support of the delegated attention, the proposed model proves its strength of maintaining and refining the essential information in both images and segmentation. It is too hard to reach this characteristic only with Convolution Neural

Network in traditional approaches.

4.3.2 Attention-driven RetinaNet

A unified network called Attention-driven RetinaNet is proposed to detect, segment, and identify the species of the microorganisms in the microscopic images. Figure 4.4 shows the architecture of the proposed network. The network architecture consists of the segmentation network, the attention block, and a detection network in a pipeline.

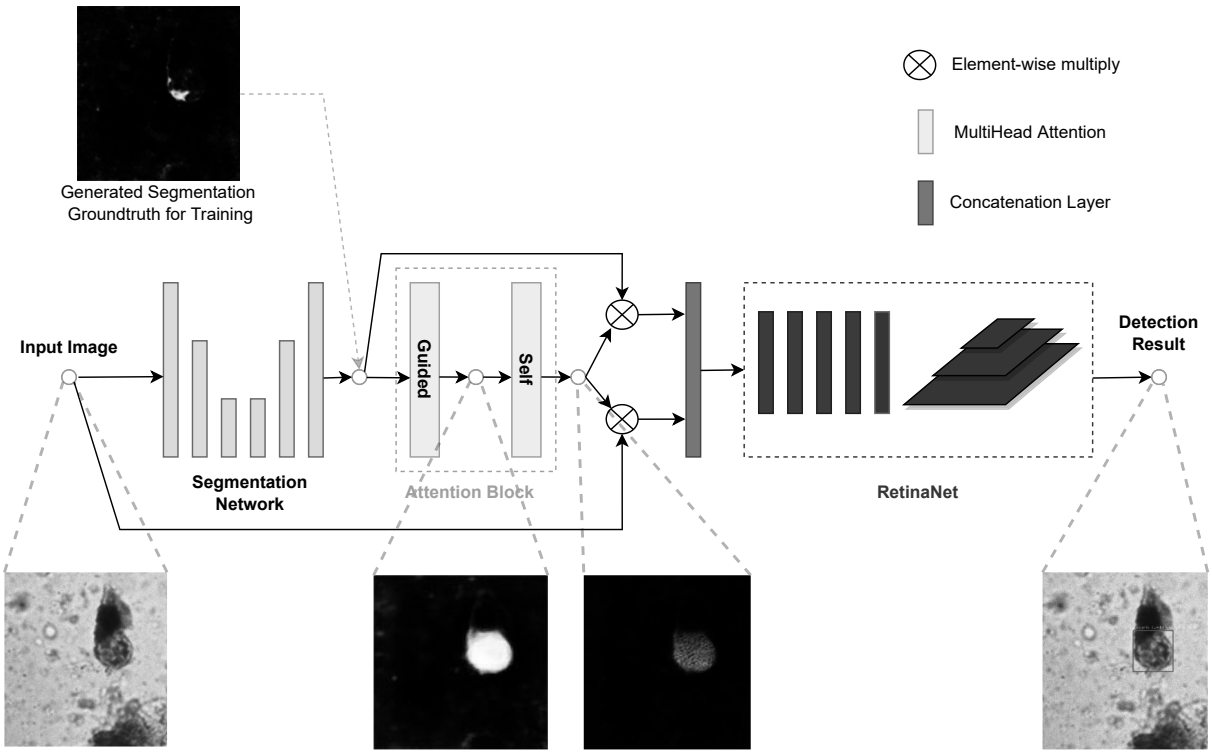


Figure 4.4: Attention-driven RetinaNet.

The proposed Attention-driven RetinaNet first produces the segmentation mask for the input image via the segmentation network. The segmentation network employs a backbone network to capture the essential information in the image and then reverse all the layers in that backbone to produce the segmentation mask. The attention block then refines the segmentation result. As shown in Figure 4.4, even though the generated segmentation ground truth for training is insufficient, the attention block still offers reasonable attention masks for the microorganisms instances. The attention mask is element-wise multiplied by the input image and the segmentation mask to filter out the background

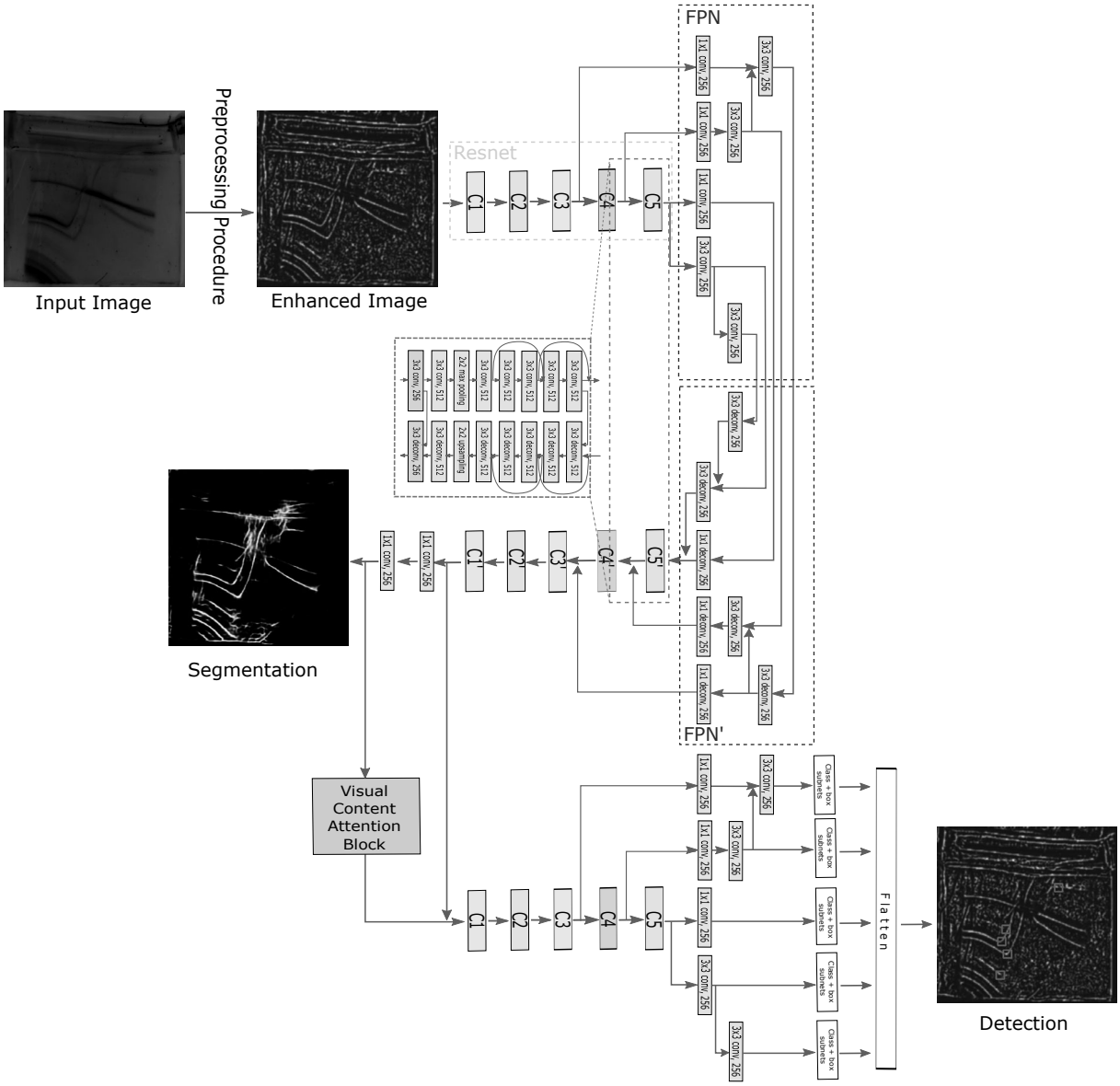


Figure 4.5: Attention-driven RetinaNet for Genome Profiling images.

and emphasize the important regions. A concatenation layer gathers all the results and feeds them to the detection network. RetinaNet is applied as the detection network to detect the target species.

The predicted segmentation result is fed into the detection network to detect the spiddos patterns. The detection network takes the segmentation mask as input and predicts the bounding boxes and their probabilities. Joining the two networks allows the detection network to adjust the weights of the segmentation network during the training phase to detect the spiddos. The quality of the segmentation results directly affects the perfor-

mance of the detection network. A connection from the C1' block of the segmentation network to the C1 block in the RetinaNet is added. This connection helps the network to learn the features that efficient not only for segmentation but also for detection.

4.3.3 Model Training

The total loss for training the unified model is as follows:

$$L = w_{Lprob}L_{prob} + w_{Lbox}L_{box} + w_{Lseg}L_{seg}, \tag{4.5}$$

where L_{prob} is the probability of bounding box prediction loss, L_{box} is the bounding box coordinate prediction loss, L_{seg} is the segmentation loss, and w_{Lprob},w_{Lbox}, w_{Lseg} are their loss-weights, respectively. In this study, the focal loss [3] and smoothed absolute value loss are applied for L_{prob} and L_{box}, respectively. For L_{seg}, the weighted cross-entropy loss function is applied. To balance the contributions of foreground and background pixels, this study applies weight for background pixel $w_b = 1$ and weight for foreground pixel

$$w_f = k\frac{\#background\_pixel}{\#foreground\_pixel} \tag{4.6}$$

which corresponds to the ratio of the number of foreground pixels to the number of background pixels, k is a hyper-parameter for controlling the contribution of foreground pixels. The segmentation loss of one image is as follows:

$$L_{seg} = -w_f \sum_{i,j,s_{i,j}=1}^{w,h} log(\hat{s}_{i,j}) - w_b \sum_{i,j,s_{i,j}=0}^{w,h} log(1-\hat{s}_{i,j}), \tag{4.7}$$

where w and h are the width and height of the image, s is the segmentation ground truth, and $\hat{s}$ is the segmentation prediction, respectively. Instead of multiple categories or multiple instances, the segmentation loss only focuses on the binary class segmentation problem where 0 and 1 correspond to background and foreground, respectively, in this study.

The transfer learning technique is applied to this problem since the number of training sample is small. A pre-trained weight of RetinaNet is applied with the Resnet50 as backbone trained on COCO dataset [82] for training the unified model. The weight is transferred for the RetinaNet instance at the initial step. In the first training phase, only the segmentation network is trained by assigning $w_{Lprob} := 0$ and $w_{Lbox} := 0$. The second training phase trains the unified network by assigning w_{Lprob} and w_{Lbox} to 1 and w_{Lseg} to 0.05, respectively. Since the priority is to detect the spiddos, w_{Lseg} should be set less than w_{Lprob} and w_{Lbox}.

4.4 Evaluation of the Attention-driven Mechanism

4.4.1 Dataset

Experiments are conducted on the TGGE image datasets. The first data set, which contains TGGE images of Bacillus coli and NIH, is divided into a training set and test set. The training set contains 8 images of Bacillus coli and 8 images of NIH. The test set contains 16 images of Bacillus coli and 16 images of NIH. The primers for references are:

- 5'-Cy3dTGCTACGTCTCTTCCGATGCTGTCTTTCGCT-3'
  5'-dTTGAATTCTATCGGTTTATCA

- 5'-Cy3-GCCG GCATCACCGGCGCCACAGGTGCGGTTG-3'
  5'-TAG CGAGGTGCCGCCGGCTTCCATTCAGGTC-3'

The number of DNA fragments varies from 5 to 6. The second data set contains 7 images of HIV. The references are

- 5'-Cy3-TGCTACGTCTCTTCCGATGCTGTCTTTCGCT-3'
  5'-TTGAATTCTATCGGTTTATCA-3'

- 5'-Cy3-GCCGG CATCACCGGCGCCACAGGTGCGGTTG-3'
  5'-TAGC GAGGTGCCGCCGGCTTCCATTCAGGTC-3'

The number of DNA fragments is 1. Samples are subjected to electrophoresis using μTGGE apparatus, micro TG, with a temperature gradient (15-65°C) set perpendicularly to the direction of DNA migration (7 min at 100V).

Table 4.1: Detection performance on mAP, precision, and recall on Bacillus coli and NIH dataset

	mAP	precision	recall
w/o Preprocessing Procedure			
Yolo	0.11	0.42	0.27
RetinaNet	0.12	0.63	0.16
Seg. RetinaNet	0.14	0.55	0.20
Att. RetinaNet	0.17	0.56	0.22
with Preprocessing Procedure			
Faster-RCNN	0.36	0.50	0.76
Mask-RCNN	0.37	0.43	0.72
Yolo	0.12	0.33	0.20
RetinaNet	0.27	0.49	0.56
Seg. RetinaNet	0.14	0.22	0.68
Att. RetinaNet	0.36	0.57	0.62

4.4.2 Results and Discussion

Genome Profiling

All the experiments are performed on a Xeon Gold 6130 2.1GHz CPU, 128 GB RAM, and an NVIDIA Tesla P100 GPU. The version of the proposed model was implemented using RetinaNet with resnet50. The batch size is set as 2 for training. The overlapping threshold for a bounding box to be matched to a ground truth box is 0.4. In the experiments, the networks are trained on the training set of Bacillus coli +NIH and are evaluated on the HIV dataset and the test set of Bacillus coli + NIH. Detection results are then compared with the ground truth spiddos only. The total number of target trajectories corresponds to the number of DNA fragments tested. Each TGGE test may contain a different number of DNA fragments, the total number of target trajectories in the images is not the same. The performance of the proposed network on spiddos detection is evaluated by the mean average precision (mAP) metric [88].

Table 4.2: Detection performance on mAP, precision, and recall on HIV dataset

	mAP	precision	recall
w/o Preprocessing Procedure			
Yolo	0.15	0.16	0.20
RetinaNet	0.10	1.00	0.07
Seg. RetinaNet	0.02	0.008	0.07
Att. RetinaNet	0.02	0.10	0.07
with Preprocessing Procedure			
Faster-RCNN	0.32	0.37	0.87
Mask-RCNN	0.24	0.21	0.80
Yolo	0.17	0.45	0.33
RetinaNet	0.12	0.17	0.80
Seg. RetinaNet	0.25	0.32	0.87
Att. RetinaNet	0.62	0.86	0.80

Table 4.3: Segmentation performance

	HIV	NIH+bacillus
w/o Preprocessing Procedure		
Seg. RetinaNet	0.98	0.97
Att. RetinaNet	0.98	0.97
with Preprocessing Procedure		
Seg. RetinaNet	0.96	0.93
Att. RetinaNet	0.97	0.94

The left panels of Fig. 4.8 and Fig. 4.9 show examples of TGGE images (in greyscale) with their ground-truth annotations. In each panel, the green rectangle indicates the area of the gradient gel in the casting chamber for the test. The red dots indicate the spiddos pattern of the image. Two of the four intersections of the white lines correspond to the flexion points of the references. In a TGGE test, the target trajectories can be obtained correctly because domain experts obtain information about which trajectories correspond to the DNA fragments. However, in the proposed method, the position of the DNA sample is unprovided; therefore, predicting the target trajectories is a challenging problem. One possible solution to overcome this issue is a relaxation that assumes all trajectories are potential trajectories. The network is designed to detect all possible spiddos. With information about DNA position, a tracing technique can be applied to remove incorrect detections; thus, the prediction accuracy of actual spiddos can be improved further. At the current stage of this study, the potential flexion points are detected as shown in the right

panels in Fig. 4.8 and Fig. 4.9. Prediction accuracy of the proposed method is compared with ground truth. Results shown in Table 4.1 depict that the precision of the proposed method achieves 0.43 which is comparable with other methods. This precision accuracy is the consequence of overestimation as assumed that any trajectories could be a potential target trajectory. The precision can be improved further by obtaining information of the positions of the DNA samples.

The detection performance of the methods on Bacillus+NIH and HIV dataset are shown in Table 4.1 and Table 4.2, respectively. For the Bacillus coli + NIH dataset, the mAP of the methods are improved by applying the preprocessing process. Except for Yolo, the recall of the methods are also improved. The magnitude of improvement in the Retina family is much higher than Yolo. It can be concluded that the proposed preprocessing procedure makes it easier for the Deep Learning-based detection networks to detect the spiddos.

Table 4.1 shows the results of the performance of the proposed method and other methods on the Bacillus coli + NIH dataset. The proposed method achieves mAP, precision, and recall values as 0.29, 0.43, and 0.61, respectively. In comparison with Seg. RetinaNet, although recall is slightly lower, the precision is higher than that of Seg. RetinaNet by almost twofold. In comparison with RetinaNet, the precision of the proposed method is lower; however, the method can provide a higher mAP value. Figure 4.8 shows example results on the Bacillus coli + NIH dataset of the proposed method. The proposed method can predict bounding boxes that are almost separated; therefore, if provided with starting points of target trajectories, target bounding boxes can be traced less ambiguously. Thus, the misdetected boxes can be eliminated more accurately. Hence, the precision of the method can be increased further. Results of the evaluation of the method and other methods on the HIV dataset are shown in Table 4.2. The proposed method can provide the mAP, precision, and recall values as 0.54, 0.71, and 0.8 respectively. Each sample in the HIV dataset contains only one DNA fragment; thus, the input images are less complicated than that of the Bacillus coli + NIH dataset. As a result, the method achieves significantly higher performance. In contrast with other methods, although the

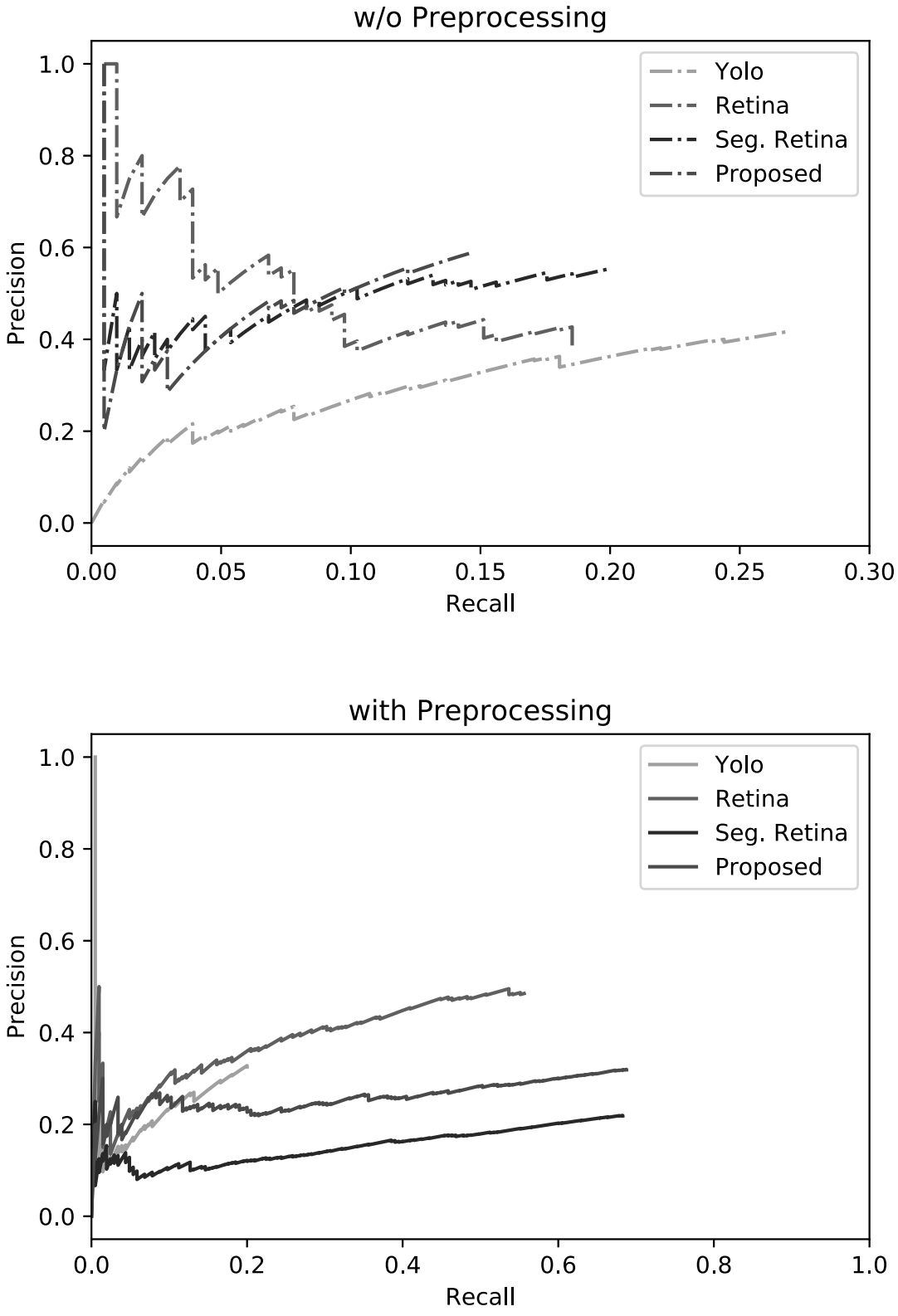


Figure 4.6: ROC curves on Bacillus coli and NIH dataset.

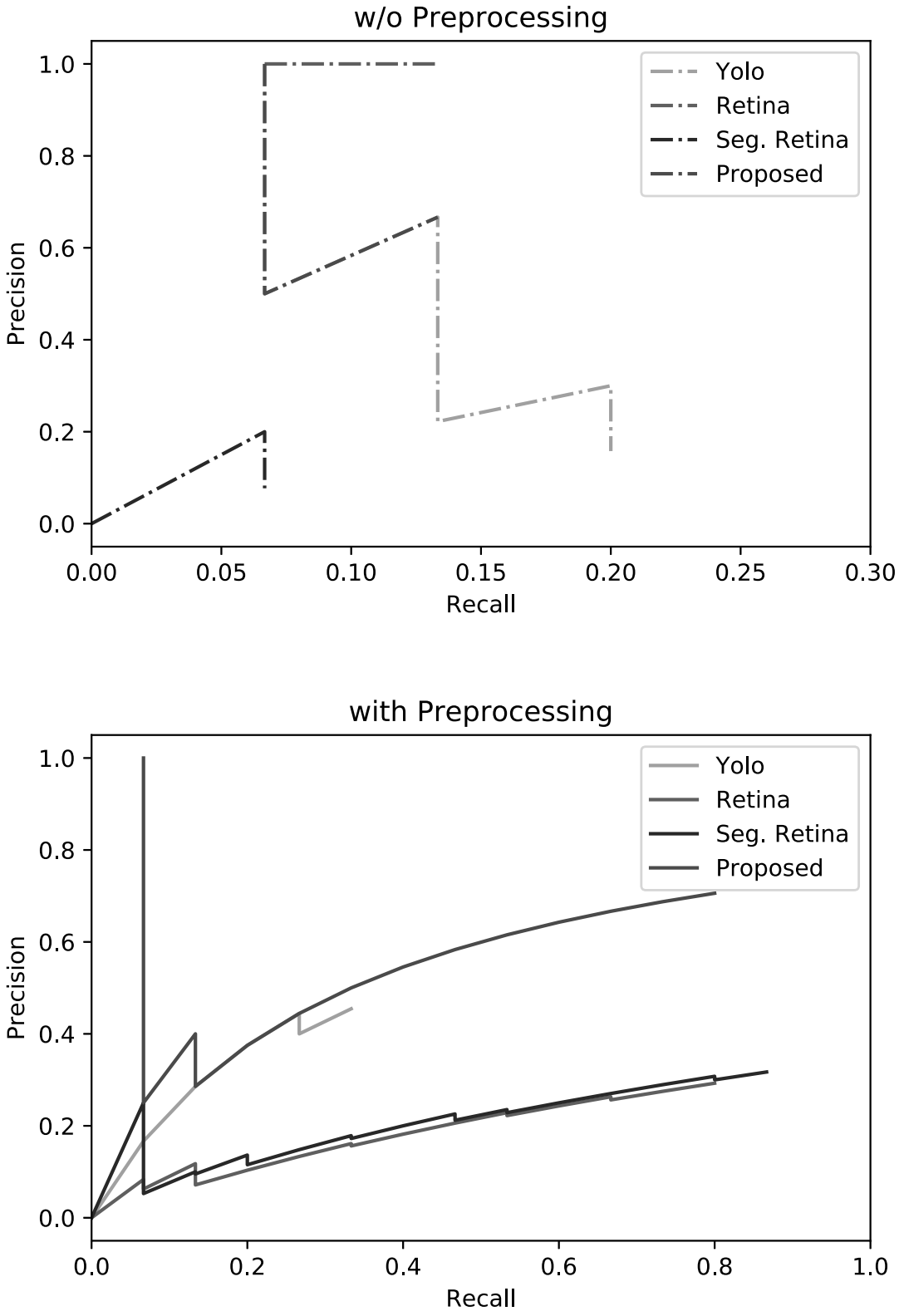


Figure 4.7: ROC curves on HIV dataset.

input images are less complicated, the performance of other methods are only slightly be improved. More precisely, the mAP and precision of the method are higher than that of Seg. RetinaNet by more than twofold. Figure 4.8 and figure 4.9 show example results on the Bacillus coli + NIH dataset and the HIV dataset of the proposed method respectively. The proposed method can predict bounding boxes that are almost separated; therefore, if provided with starting points of target trajectories, target bounding boxes can be traced less ambiguously. Thus, the misdetected boxes can be eliminated more accurately. Hence, the precision of the method can be increased further.

Precision-recall curves show the trade-off between the true positive and the predictive value for a predictive model using different probability thresholds. The detector predicts the probabilities of the target class to provide the capability to choose and calibrate the threshold toward the trade-off between precision and recall. Reviewing both precision and recall is useful where there is an imbalance in the observation between two classes. In case of detection, two classes are the target class and the background class. All the predictions are sorted with respect to the probabilities in descending order. The precision and recall are calculated after every single prediction. The tested method starts at (0,1) if the prediction with the highest probability is correct and at (0,0) otherwise. The precision-recall curves of the networks on Bacillus coli + NIH and HIV are shown in the Fig. 4.6 and Fig. 4.7, respectively. The curves drop rapidly at the beginning means that the predicted spiddos with the highest confidence are misdetections. This phenomenon commonly occurs due to the limited number of training data in the method that leads to the uncertainty of the feature of the target class. In this case, there are similar appearances between the spiddos and uninterested flexion points or noise dots. The curves then rise steadily to the end means that the predicted spiddos with lower confidences are accurate. The correct predicted spiddos with lower confidence dues to the small number of training samples and the relaxation. For the HIV dataset, the proposed method achieves the highest precision value at each specific level of recall. It shows that the proposed method outperformed the other methods.

Table. 4.3 shows the segmentation performance of Seg. RetinaNet and the proposed

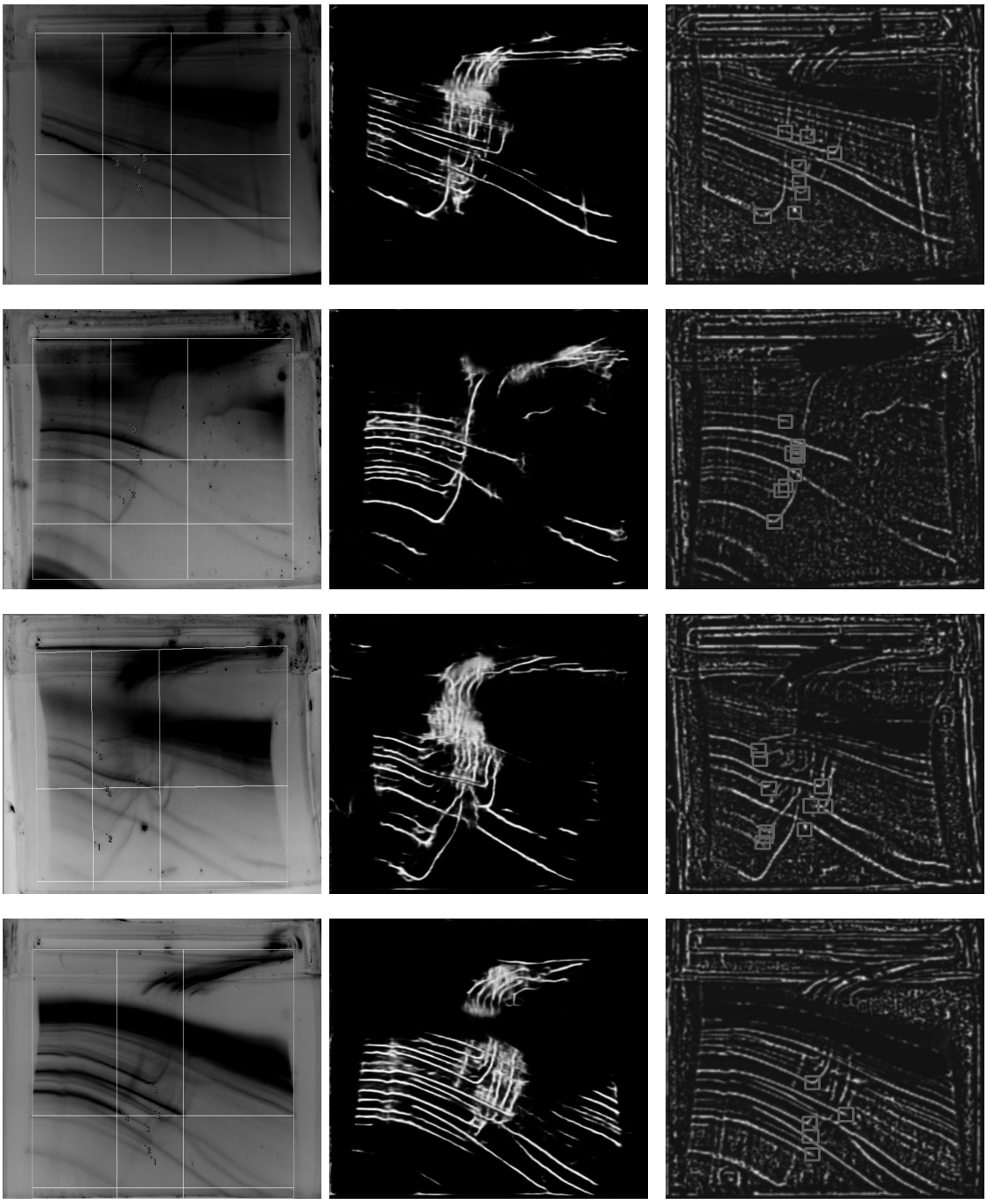


Figure 4.8: Examples of the results of the Attention-driven RetinaNet on Bacillus coli + NIH dataset. The left column is the input image in grayscale and ground-truth annotated in color. The red dots indicate the ground-truth spiddos. The middle column is the segmentation results and the right column is the results.

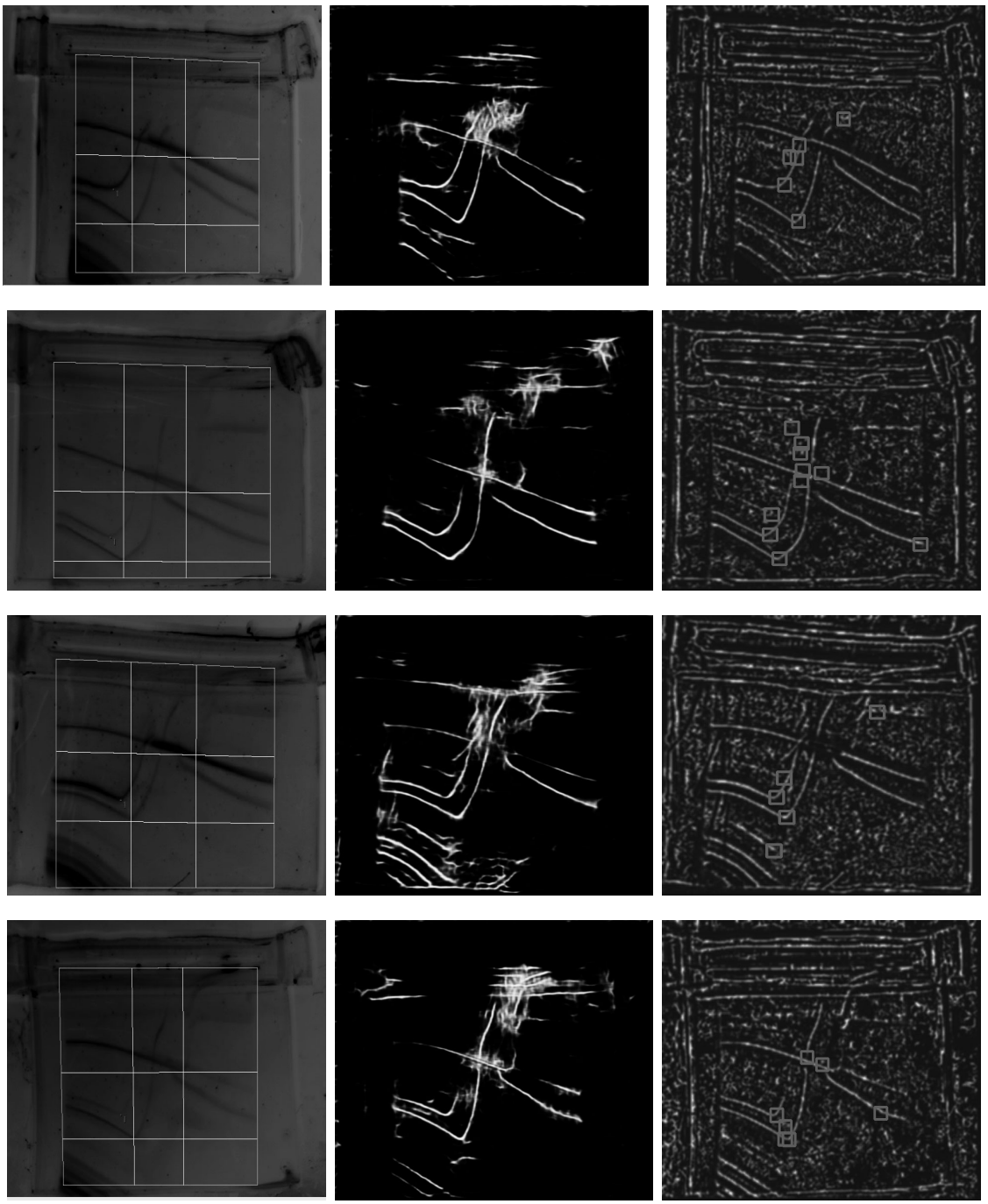


Figure 4.9: Examples of the results of the Attention-driven RetinaNet on HIV dataset. The left column is the input image and the ground-truth. The middle column is the segmentation results and the right column is the results.

network. The proposed network achieves the equivalent results to the Seg. RetinaNet. Both Seg. RetinaNet and the proposed method segmentation accuracies are slightly reduced when applying the preprocessing procedure. An explanation is that the trajectories get shaper and thinner after the preprocessing procedure. The middle panels in Fig. 4.8 and Fig. 4.9 show examples of segmentation results of the proposed method. The proposed method is able to segment the trajectories in the clear regions and predict the trajectories in occluded regions. If the occluded regions contain multiple trajectories, the segmentation result is blurred in those regions. This study only performs semantic segmentation that distinguishes pixels of the trajectories between that of the background. Instance level segmentation that indicates the pixels of different trajectories may help to remove the misdetections. This is left for future works.

In this study, the network is trained with a small number of training samples. Since the GP method is on the way to be established, researchers are investigating various directions to improve. The choices of referenced DNA fragments are also under investigation. Moreover, the TGGE is labor-intensive in nature. Therefore, the number of samples using the same referenced DNA fragments is limited. However, the proposed method is not limited to the choices of the referenced DNA fragments. The precision value can be improved further when there are more training data.

Protozoa

For protozoa domain, experiments are conducted on the ICIP 2022 Grand Challenge Parasitic Eggs dataset which consists of 11000 microscopic images of 11 categories. The dataset is divided into the training set and test set. The training set consists of the 800 samples of each category; therefore, 8800 images in total.

The size of the input for MultiHead Attention is relatively small compared to the input images. The input image need to be resized before feeding it into the MultiHead Attention layer and resize the attention output to the input image. With a more powerful GPU, increasing the size of the inputs for MultiHead Attention is expected to yield better results.

Table 4.4: mAP on ICIP 2022 protozoa dataset.

mAP	ω_0	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	ω_7	ω_8	ω_9	ω_{10}	**ave.**
DeTR [93]	0.09	0.76	0.18	0.09	0.18	0.00	0.09	0.25	0.09	0.09	0.18	0.18
Faster-RCNN [2]	0.17	0.59	0.42	0.18	0.31	0.09	0.09	0.23	0.72	0.52	0.27	0.33
RetinaNet	0.72	0.72	0.91	0.81	0.91	0.27	0.91	0.91	0.82	0.91	0.91	0.80
Seg-RetinaNet	0.10	0.18	0.07	0.16	0.18	0.09	0.18	0.40	0.09	0.33	0.07	0.17
Att-RetinaNet	0.82	0.88	0.90	0.82	0.91	0.18	0.90	0.91	0.91	0.91	0.91	**0.82**
Recall	ω_0	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	ω_7	ω_8	ω_9	ω_{10}	**ave.**
DeTR	0.01	0.89	0.17	0.05	0.11	0.00	0.02	0.45	0.04	0.10	0.13	0.18
Faster-RCNN	0.16	0.77	0.45	0.12	0.40	0.02	0.09	0.23	0.75	0.51	0.24	0.34
RetinaNet	0.70	0.80	0.92	0.87	0.91	0.29	0.92	0.92	0.84	0.99	0.99	0.83
Seg-RetinaNet	0.05	0.15	0.04	0.11	0.12	0.02	0.14	0.44	0.01	0.31	0.04	0.13
Att-RetinaNet	0.89	0.94	0.96	0.89	0.97	0.16	0.93	0.95	0.94	0.98	0.97	**0.87**

Table 4.5: mAP of few shots learning.

mAP	ω_0	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	ω_7	ω_8	ω_9	ω_{10}	**ave.**
DeTR	0.18	0.18	0.18	0.18	0.27	0.09	0.18	0.18	0.18	0.27	0.18	0.19
Faster-RCNN	0.18	0.17	0.17	0.18	0.27	0.18	0.18	0.18	0.18	0.27	0.27	0.20
RetinaNet	0.16	0.26	0.11	0.13	0.16	0.33	0.15	0.03	0.15	0.11	0.15	0.16
Seg-RetinaNet	0.07	0.17	0.07	0.14	0.17	0.27	0.17	0.30	0.08	0.25	0.09	0.16
Att-RetinaNet	0.17	0.14	0.14	0.14	0.09	0.44	0.17	0.17	0.20	0.17	0.25	0.19

Table 4.4 shows the detection results of all methods. The categories $\omega_0, ..., \omega_{10}$ correspond to the category list of the Grand Challenge, i.e., *Ascaris lumbricoides*, *Capillaria philippinensis*, *Enterobius vermicularis*, *Fasciolopsis buski*, *Hookworm egg*, *Hymenolepis diminuta*, *Hymenolepis nana*, *Opisthorchis viverrine*, *Paragonimus spp*, *Taenia spp. egg*, and *Trichuris trichiura*, respectively. The Att-RetinaNet achieves the highest mAP value of 0.82 on average. The network only achieves 0.18 in mAP on the ω_5, which is *Hymenolepis diminuta*, due to the low recall (0.16). Without quality segmentation ground truth, Seg-RetinaNet [94] fails to segment and therefore fails to detect the microorganisms.

Figure 4.10 shows the segmentation, first and second attention results of the corresponding input images. Segmentation masks mark all the target objects' instances in the images. It is easy to cover the objects' regions in the segmentation masks. The first attention, which is the guided multi-head attention, refines the boundaries of the object instances in the segmentation. While the inner regions of the objects' instance are weighted equally in the first attention, the second attention, which is the self multi-head attention, weights the essential parts for the identification task. In the results, it can be

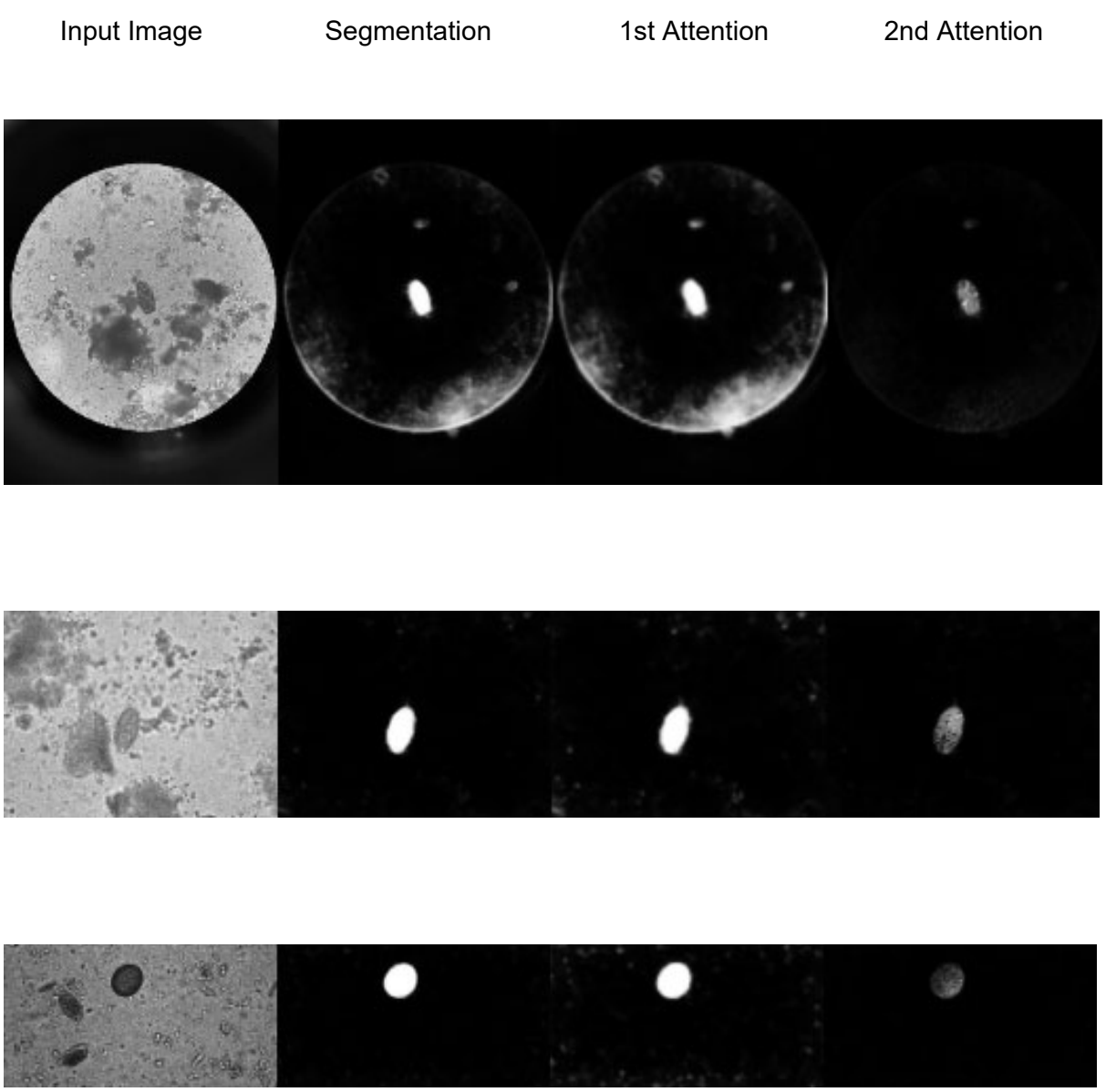


Figure 4.10: Segmentation and Attention Results of Attention-driven RetinaNet on Protozoa image.

interpreted that the second attention focuses more on one end of the protozoa instances. Discriminately focusing on the objects' instances leads to the improvement of detection.

Researches on the automatical detection of microorganisms often suffer from the small number of training data problems. Therefore, it is useful to build a network that can be trained with few labeled samples. It is difficult to collect a large number of microorganisms with a broad set of species. Segmentation-driven RetinaNet is originally trained on at most 5 samples per category but still offers promising results. In this experiment, the networks' ability on a few-shot learning problem is tested.

The 5 samples of each category in the ICIP 2022 Grand Challenge Parasitic Eggs dataset are collected for the training set. Since the microorganisms appear at arbitrary orientations, the images are rotated to generate more data without creating unrealistic samples. The training set is augmented by rotating images at the center of the target object's bounding box with an interval of 5 degrees. The rest of the dataset, which is 10945 images, is reserved for testing.

Table 4.5 shows the results of the few-shot learning problem. Faster-RCNN achieves the highest mAP (0.20) in this scenario with 0.99 precision. Our proposed Att-RetinaNet achieves slightly higher mAP (0.19) than others in the RetinaNet family (0.16). This result also shows the potential of our proposed network in applying it to real-world applications where collecting data is challenging.

4.5 Summary of Finding distinctive features

This chapter focuses on finding the distinctive features for objects with less visual information to enhance identification accuracy. The segmentation-driven mechanism is the key to guiding the deep network in finding the characteristic features for detection. However, the segmentation results produced by deep networks are coarse and blurred at the boundaries of the objects. The segmentation-driven mechanism is replaced with the attention-driven mechanism to find the distinctive features. Moreover, the attention mechanism is able to refine the mistake in segmentation results.

The proposed Attention-driven mechanism is applied to guide the network on the characteristic and distinctive features. The attention-driven mechanism is also applied to the protozoa domain. By predicting the segmentation, the network can keep track of the trajectories even in the occluded, unclear regions. The attention mechanism is applied to enhance the segmentation mask. The network then detects the spiddos patterns in the attention mask. For the Genome Profiling domain, the proposed Attention-driven RetinaNet is trained on 16 samples of the dataset containing Bacillus coli and NIH genomes. The network is evaluated on the test set containing 32 Bacillus coli + NIH and 7 samples of the HIV dataset. Even though there are few samples for training, the network can still detect the spiddos and achieves mAP of 0.29 and 0.54 on the two datasets. The proposed network outperforms other related methods with respect to the mAP metric. The results show that the proposed method can predict the segmentation of the trajectories even in unclear and overlapping regions. Experiments are conducted on the extensive training set and few-shot learning scenarios for the protozoa domain. The network achieves the highest mAP in both cases.

Chapter 5

Polymorphism

5.1 Polymorphism of a Category

In many domains, there are multiple appearances in the same class. This problem is called the polymorphism. The multiple appearances in polymorphism problems may come from different species, living conditions, evolution, etc. Each appearance contains a smaller set of the feature set of its class. For example, there are several types of cats. They share similar appearances since all of them are cats. Different poses of an object are also examples of multiple appearance problems. Their appearance may also roughly change during their lives.

There are several attempts that try to solve the polymorphism problems in object detection. DPM [63] divides the samples of a class into multiple poses concerning the aspect ratios. For example, the bounding boxes of cars in the front view have the shapes of squares, while the side view has the forms of horizontal rectangles. The details of the front view are also much different from the details of the side view. Clustering the target object samples into groups of aspect ratios and training the detectors for each group independently improve the detection performance. The idea is then derived from the anchor box mechanism in modern detection networks. In detection neural networks such as Yolo [81], SSD [78], or RetinaNet [3], the networks actually make thousands of predictions, and only the those considered as objects are shown. A dense grid of bouding

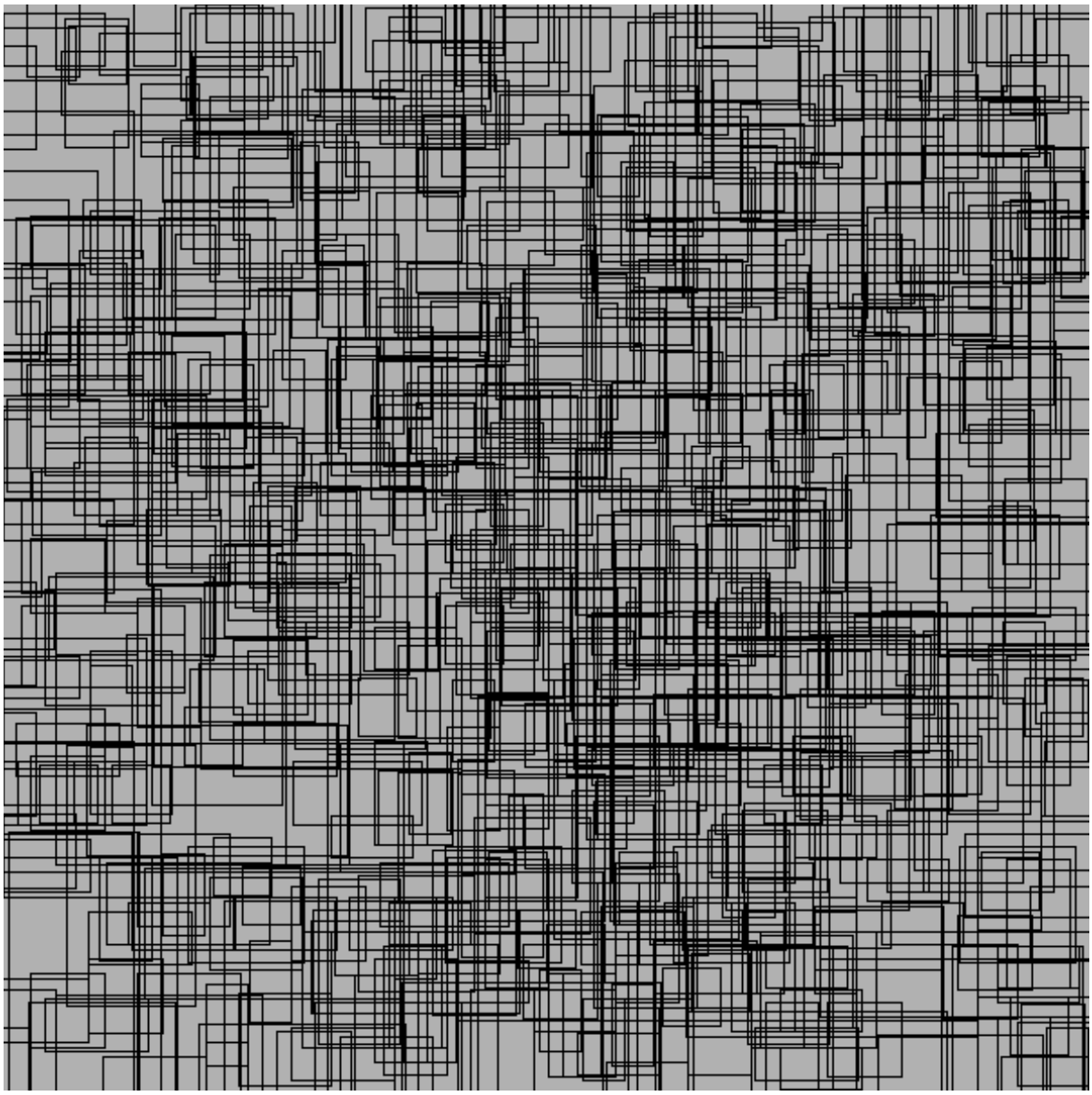


Figure 5.1: Example of the grid of anchor boxes in RetinaNet [3] that applied on the input image to detection multiple object with different sizes and aspect ratios.

boxes is applied to the input image to detect the object at multiple locations. Each cell of the grid contains several classifiers that are in charge of several image regions with different sizes and aspect ratios. Figure 5.1 shows an example of anchor boxes in RetinaNet. The classifier detects the target object instance at the closest cell on the grid with the closest size and aspect ratio. However, there are differences among the intraclass subcategories, while there are similarities among the interclass subcategories. For example, there are objects such as a face in which different poses have the same bounding box aspect ratios. Therefore, a more sophisticated criteria is required to cluster the training samples.

Due to the regions and living conditions, species may change their appearances. One possible solution is to collect many samples of all the appearances and train with robust classifiers. However, this approach is insufficient in the domain that is difficult to collect data. Few numbers of samples for each appearance lead the detector to overfit that appearance.

In biology, experts cluster the living organisms into a taxonomy based on their appearances, structure of the organs, living habitats, etc. A living organism is clustered into the kingdom, phylum (or division), class, order, family, genus, and species. With the taxonomy relationship, similar species are separated. Hierarchical classifiers have been widely applied to improve classification performance. In general daily life objects, there is also a hierarchical relationship. Levatic et al. [95] discuss the importance of the label hierarchy on the classification tasks. Several hierarchical classifier based methods have been developed such as SVM [96; 97], Decision Tree [98; 99], Artificial Neural Network [100], etc. For object detection, Fan [101] proposed a method that clusters similar categories into a new category to build a hierarchical relationship of image patches for localization tasks. Bueno et al. [102] proposed a detection method with Deep Reinforcement Learning that considers the hierarchical relationship of image regions.

The detection accuracy is often improved by detecting the dis-ambiguous sub-categories. However, the information on the relationship between the sub-categories which belong to the same category is not considered. Flat classifiers only find distinctive features of the target classes. Multiple appearances of a class confuse the detector during training phase.

A feature of one appearance may not appear in others. Those flat classifiers tend to focus on a small group of differences enough to classify the target classes.

A hierarchical multi-label tree is applied to perform the hierarchical classification. Leave-one-out strategy, which trains an independent classifier at every node in the hierarchical relationship tree, is applied. At each node, the classifier is trained for the binary classification task of that category versus the rest. This study employs a hierarchical classifier for multi-label problems. The hierarchy for RetinaNet is proposed to be integrated on the natural hierarchical relationship of protozoa species and their life-cycle stages.

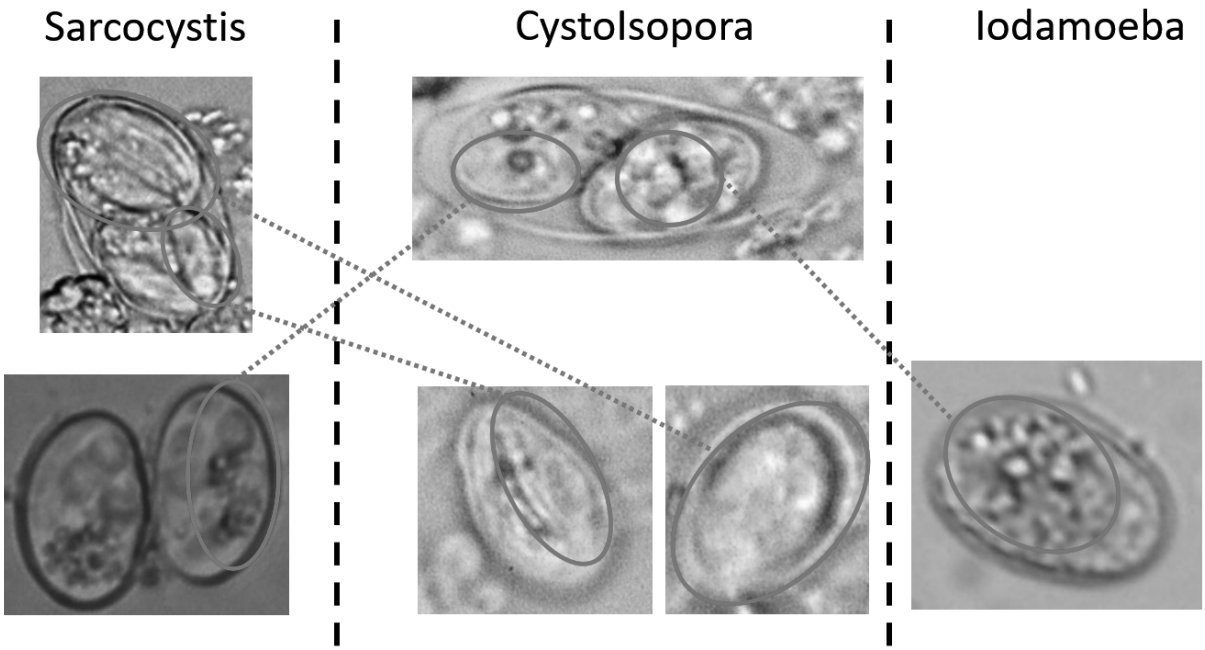


Figure 5.2: Examples of protozoa that have similar appearances.

Different appearances of a category are treated as sub-categories. This chapter explores the relationship between the sub-categories in the protozoa domain. There are a lot of variations in the appearances of protozoa during their life-cycle stages. On the other hand, a life-cycle stage of a species may have similar appearances to the life-cycle stages of different species. Figure 5.2 shows examples of protozoa that have similar appearances. The double wall-celled stage of *Sarcocystis* has two cysts inside a thick cell wall that is similar to *Cystoisospora*. Sporulated oocyst stage of *Sarcocystis* is a single thick cell wall

that is similar to *Cyclospora* and *Iodamoba*. Since there are a lot of differences among stages, even for one species, it is hard to find general distinctive features to represent each species.

The main contributions of this study in this problem are as follows:

- Exloring the hierarchical relationship of the target categories. In case of protozoa domain, samples from each category of protozoa species are clustered with respect to their life-cycle stage.
- Integrating the hierarchical classifier to deep network detector to form a end-to-end network that detect the target object with its hierarchical structures simultaneously. This study proposed Segmentation-driven Hierachical RetinaNet to detect and identify the protozoa with hierarchical multi-label.

5.2 Segmentation-driven Hierarchical RetinaNet

5.2.1 Hierarchical Relationship in Protozoa domain

Even though there is a small number of training samples, the hierarchical classifier is employed to improve the performance when training samples appear differently. This study employs the hierarchical relationship in the protozoa domain. The protozoa often have two main stages: cyst and trophozoites. Each species' life cycle is divided into multiple stages that differ in morphology, such as the number of oocysts, wall cells, and internal nuclei. Dividing each species into multiple life-cycle stages reduces the difficulty of the generalization tasks of the model. It also helps clarify the distinctive features for classification between similar stages across the species.

The network can be trained to predict the probabilities of the life-cycle stages instead of the species. The predicted life-cycle stage, which has the maximum score, is then mapped to the corresponding species category. Generally, predicting sub-categories yields better results than predicting categories directly. Compared to the goal of detecting categories,

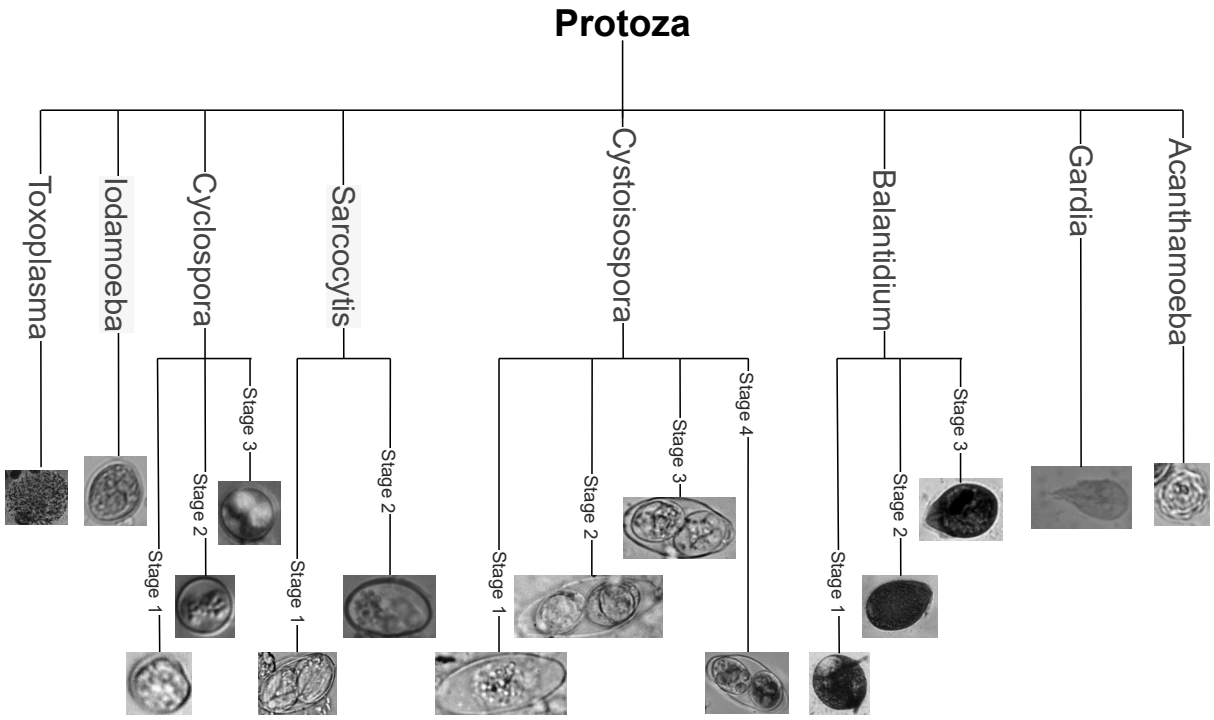


Figure 5.3: The hierarchical relationship tree for the protozoa dataset.

detecting the subcategory forces the network to learn more efficient features to distinguish the similar stages of different species.

Identifying the subcategories independently only focuses on the unique features of each subcategory. The similarities between intra-class subcategories are not considered in this manner. The hierarchical classifier is applied instead of the original flat classifier to reveal the similarities of intraclass subcategories. The relationship between the subcategories for life-cycle stages and the categories for species is considered as a hierarchical relationship. Generally, the flat classifiers predict the subcategories and then infer the corresponding category. On the other hand, the hierarchical classifier predicts a pair of subcategories and their category. The hierarchical relationship is implicitly captured in the predictions. Figure 5.3 shows the hierarchical relationship in the protozoa dataset.

5.2.2 Segmentation-driven Hierarchical RetinaNet Architecture

This chapter integrates a hierarchical classifier into the Segmentation-driven RetinaNet for the protozoa domain. The flat classifier of the detection network instance is replaced

with the hierarchical classifier. In the architecture of the detection network, multiple classification sub-networks are organized in a grid after the Feature Pyramid Network to predict the categories of the bounding boxes at numerous locations in the input image. Each classification sub-network is a flat classifier that predicts the probabilities of either the categories or the sub-categories. In the hierarchical classification, these sub-networks predict the probabilities of the life-cycle stages. These layers are called sub-category classification layers. More fully connected layers are added after the sub-category classification layers of these sub-networks to predict the probabilities of the species. These sub-category and category classification layers organize the hierarchical classifiers that predict pairs of the life-cycle stages and the species. The focal loss is applied for sub-category and category layers during the training phase.

In this protozoa problem, each sub-category corresponds to exactly one category. There is a trivial case in which the network only weights the corresponding sub-category nodes in the case of predicting the probability of a particular category. A residual connection is added from the Feature Pyramid Network to the category classification layers to prevent the network from learning the trivial case. This residual connection helps the network obtain the result of the sub-category classification layers and the extracted features of the Feature Pyramid Network. Moreover, the residual connection forces the network to learn the shared features for grouping the sub-categories into the corresponding categories. The extracted features are then shared to identify both sub-categories and categories. Therefore, the benefit of the hierarchical classifier is that the network is forced to simultaneously learn the distinctive features among the sub-categories and the similarities between intraclass sub-categories.

The architecture of the hierarchical classifiers for Segmentation-driven RetinaNet is illustrated in Fig. 5.4 where A is the number of anchor boxes in RetinaNet, K is the sum of all the life-cycle stages, and S is the number of target species. The orange layer and the purple layers are the subcategory classification layer and the category classification layer, respectively. The architecture is able to be extended to more layers of hierarchy when there are more layers in the hierarchical relationship structure. In this chapter,

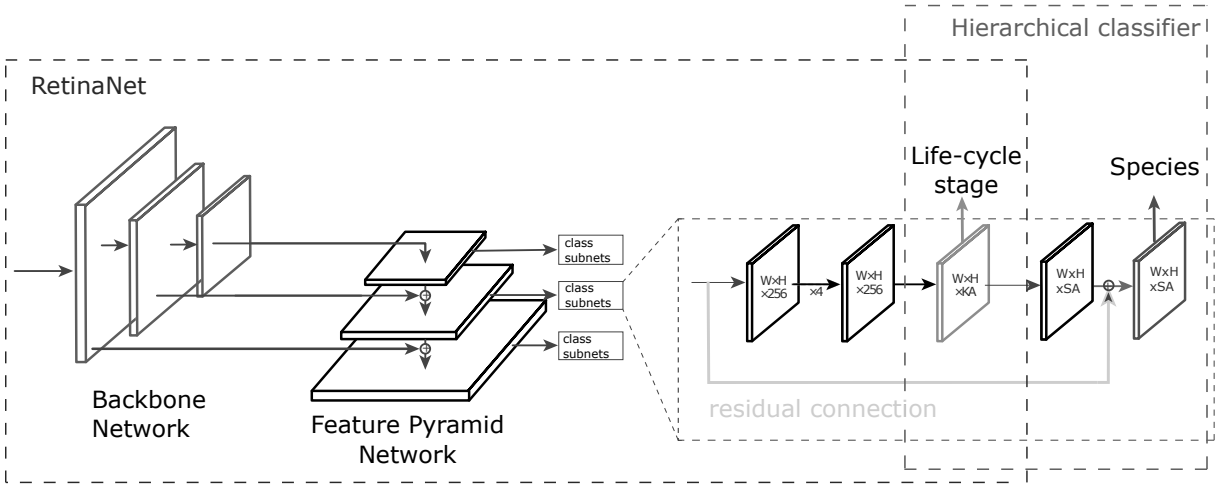


Figure 5.4: Hierarchical classifier for RetinaNet. For protozoa domain, the subcategory classifier corresponds to the life-cycle stage while the category classifier correspond to the species.

there are only two layers in the hierarchy: category and its first level of subcategory for the protozoa domain.

5.3 Evaluation

The Segmentation-driven Hierarchical RetinaNet is evaluated on the protozoa dataset. The results of the original RetinaNet, the sub-network used for training, Segmentation-driven RetinaNet (Seg. Retina) and the Segmentation-driven Hierarchical RetinaNet (Seg. Hier. Retina) in both cases with and without the sub-category of the life-cycle stages are shown in Table 5.1.

The Seg. Hier. Retina achieves mAP, precision, and recall values on the average of 0.82, 0.85, and 0.94, respectively. Those are the highest results on among the tested methods. By applying the hierarchical classifier, the detection performance increases 5% in mAP, 7% in precision, and 4% in recall, respectively. The hierarchical classifier improves the performance of the network by reducing the overlapped bounding boxes (for example, the case of *Aca* and *Cca*). Figure 5.5 shows the example segmentation and detection results of the Seg. Hier. Retina. However, the scores of the predicted classes of Seg. Hier. Retina is lower than Seg. Retina in these samples (Figs. 3.9 and 5.5). Seg. Hier. Retina misses

Table 5.1: Detection performances on mAP, precision, and recall with respect to species.

	Cca	Sar	Aca	Ibu	Tgo	Gla	Bco	Cbe	avg.
mAP									
On category									
RetinaNet	0.63	0.42	0.17	0.12	0.0	0.14	0.54	0.00	0.25
Sub-network	0.86	0.00	0.34	0.61	0.61	0.45	0.53	0.10	0.44
Seg. Retina	0.86	0.25	0.36	0.67	1.00	1.00	0.58	0.30	0.65
On sub-category									
RetinaNet	0.82	0.10	0.53	0.18	0.00	0.36	0.39	0.00	0.30
Sub-network	0.86	0.18	0.78	0.73	1.00	1.00	1.00	0.32	0.73
Seg. Retina	0.89	0.50	0.73	0.84	1.00	0.84	0.61	0.75	0.77
Seg. Hier. Retina	0.77	0.42	0.77	0.96	1.00	1.00	0.90	0.75	0.82
Precision									
On category									
RetinaNet	0.69	0.40	0.19	0.67	0.00	0.22	0.50	0.00	0.33
Sub-network	0.95	0.00	0.42	0.75	0.67	0.50	0.58	0.14	0.50
Seg. Retina	0.94	0.33	0.44	0.82	1.0	1.0	0.64	0.33	0.69
On sub-category									
RetinaNet	1.00	0.14	0.48	1.00	0.00	0.50	0.50	0.00	0.45
Sub-network	0.94	0.25	0.74	1.00	1.00	1.00	1.00	0.50	0.80
Seg. Retina	0.97	0.50	0.71	0.90	1.00	0.75	0.83	0.60	0.78
Seg. Hier. Retina	0.94	0.67	0.77	0.91	1.00	1.00	0.87	0.60	0.85
Recall									
On category									
RetinaNet	0.94	0.67	0.94	0.20	0.00	0.67	1.00	0.00	0.55
Sub-network	0.97	0.00	0.88	0.90	1.00	1.00	1.00	0.67	0.80
Seg. Retina	0.94	0.67	0.88	0.90	1.00	1.00	1.00	1.00	0.92
On sub-category									
RetinaNet	0.83	0.67	0.94	0.10	0.00	0.33	0.43	0.00	0.41
Sub-network	0.94	0.67	1.0	0.70	1.00	1.00	1.00	0.67	0.87
Seg. Retina	0.92	0.67	1.00	0.90	1.00	1.00	0.71	1.00	0.90
Seg. Hier. Retina	0.83	0.67	1.00	1.00	1.00	1.00	1.00	1.00	0.94

few instances of *Cca*; therefore, the recall on *Cca* of Seg. Hier. Retina is lower than Seg. Retina.

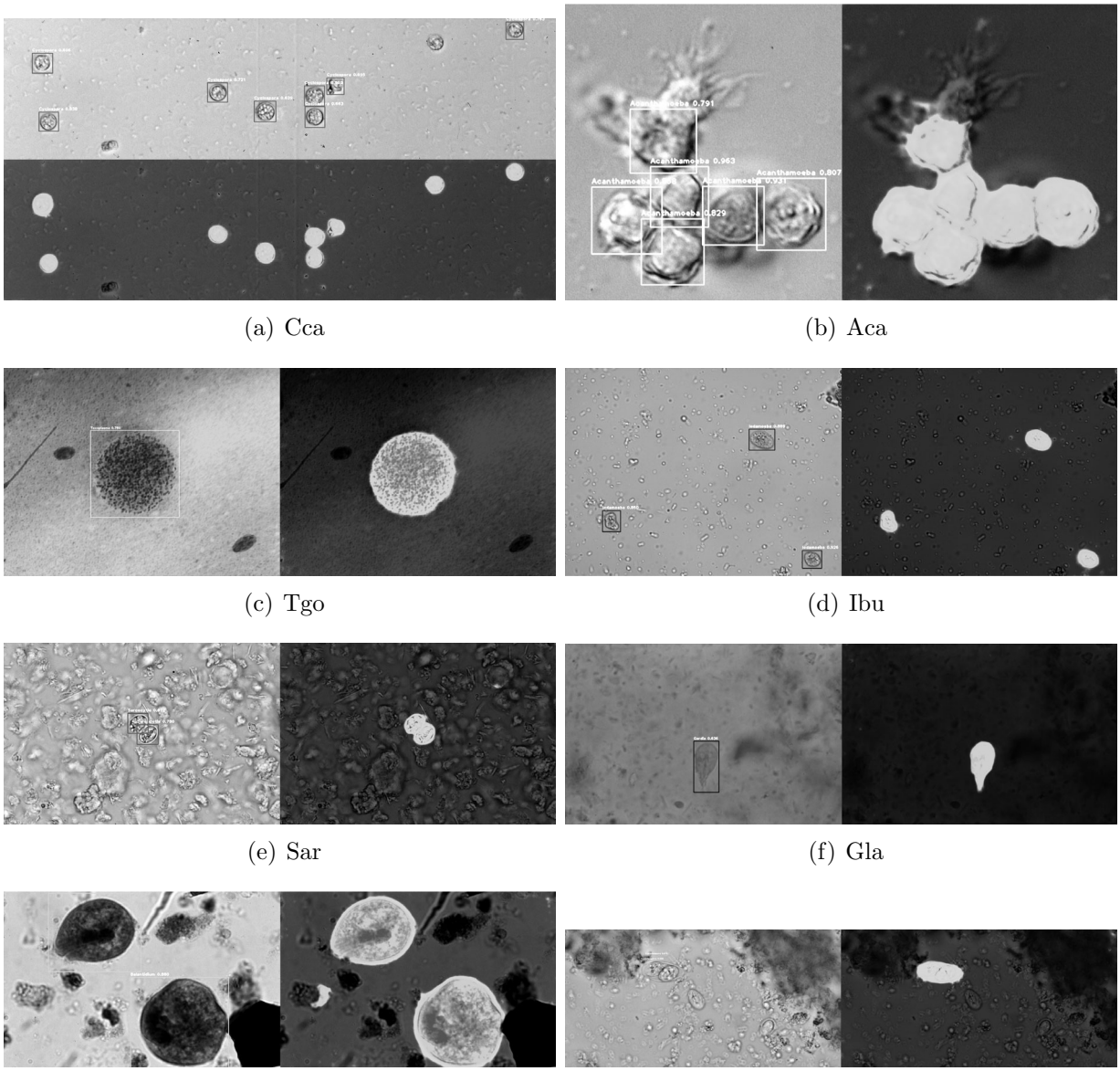
Table 5.2 shows the segmentation results of the Seg. Retina and Seg. Hier. Retina. The Seg. Hier. Retina achieves average binary accuracy, precision, and recall of 0.98,

Table 5.2: Segmentation performance.

	Cca	Sar	Aca	Ibu	Tgo	Gla	Bco	Cbe	avg.
Binary Accuracy									
Seg. Retina	0.99	0.97	0.97	1.00	0.99	0.99	0.86	0.98	0.96
Seg. Hier. Retina	0.99	0.92	0.96	1.00	0.99	0.98	0.99	0.99	0.98
Precision									
Seg. Retina	0.84	0.74	0.90	0.82	0.91	0.76	0.96	0.55	0.86
Seg. Hier. Retina	0.86	0.89	0.90	0.83	0.91	0.87	0.98	0.75	0.93
Recall									
Seg. Retina	0.99	0.96	0.96	0.99	1.00	0.94	0.85	0.95	0.91
Seg. Hier. Retina	0.99	0.95	0.94	1.00	1.00	0.89	0.92	0.94	0.95

0.93, and 0.95, respectively. The segmentation performance is slightly improved in binary accuracy, precision, and recall. Since the network is a built-in end-to-end pipeline, the detection network modifies the segmentation result to get higher detection performance. Therefore, the segmentation performance is improved when the detection performance is improved.

Figure 5.6 shows an example of detection result with probabilities of the categories of Seg. Retina and Seg. Hier. Retina. In this example, Seg. Hier. Retina outperforms Seg. Retina when successfully detects and identifies all the protozoa instances. The Seg. Retina with the flat classifier tries to maximize the probabilities of the categories with similar parts to the sample. This is because the flat classifiers learn the efficient features to distinguish among the target category set. The appearances in the training samples cannot cover all of those features in the case of a small number of training data. Therefore, the network reffers them as distinctive features for other categories. In the example, the probabilities of *Sar* and *Iod* produced by Seg. Retina is high, which leads to the misidentified instance. On the other hand, Seg. Hier. Retina allows the category classifier layer to adjust the probabilities produced by the subcategory classifier layer. In the example, Seg. Hier. Retina adjusts the probability of *Sar* lower and that of *Iod* higher.



(a) Cca (b) Aca

(c) Tgo (d) Ibu

(e) Sar (f) Gla

(g) Bco (h) Cbe

Figure 5.5: Examples of detection and segmentation results of Segmentation-driven Hierarchical RetinaNet.

5.4 Summary of Polymorphism

This study discusses the polymorphism of objects with less visual information. This study explores the relationship of various appearances of the target categories. This study introduces Segmentation-driven Hierarchical RetinaNet, which integrate the hierarchical classifier to a detection network for the various appearance problem for the object with less visual information. By applying the hierarchical structure and clustering the different

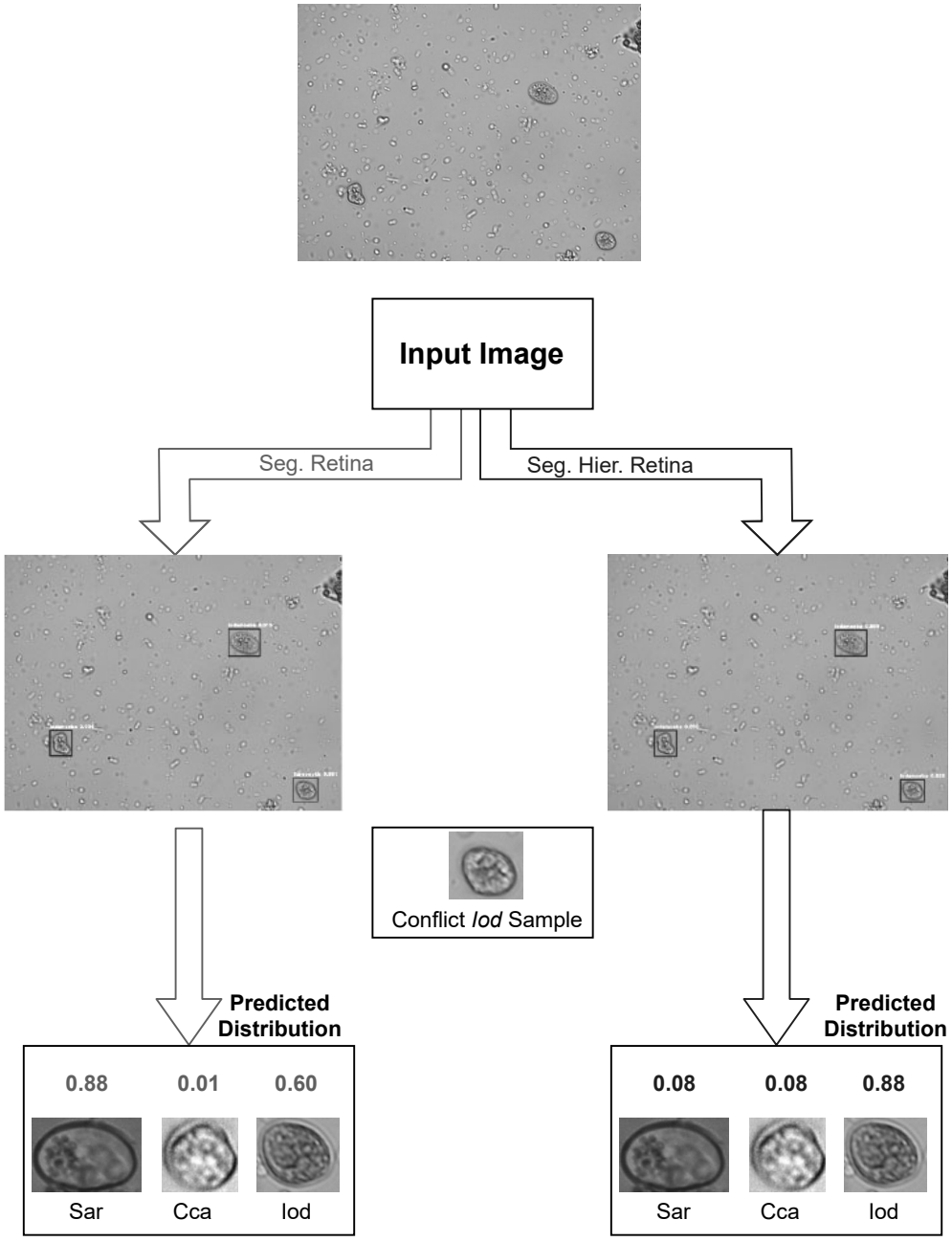


Figure 5.6: Category probabilities predicted on an example input by Seg. RetinaNet and Seg. Hier. RetinaNet.

appearances, the ambiguity in each category is cleared. The Hierarchical Segmentation-driven RetinaNet is evaluated on the protozoa dataset. Experimental results show that the proposed network can learn the characteristic features among similar target objects.

Chapter 6

Conclusions and Future Work

This chapter summarizes the key contributions and the main observation obtained from the experiments. This chapter also points out some potential research directions for future work on detection for objects with less visual information.

6.1 Conclusion

This study is motivated by the need for a comprehensive method for medical image domains to accelerate the diagnosis and treatment processes. This dissertation concerns the detection, segmentation, and identification tasks for object with less visual information. The target objects of this dissertation are objects that:

- have less visual information: decomposing the structure of the target objects yields shallower part-based hierarchical structures than daily life objects.
- have a small number of training data: target objects in the domains that are difficult to collect.
- have various appearances: each appearance shows a smaller number of features than that of the feature set of its category.

The research question of this dissertation is: "How can we identify a category with a small number of visual information?". Finding characteristic and distinctive features is

the essential to detect and identify the target categories. Those are shape and texture features. This study also considers the distinctive features among similar appearance objects.

To answer the research question, the dissertation focuses on:

- Finding characteristic features: This study first finds the texture features for the target objects. Those features characterize the appearance of the target objects. In the case of the small dataset, neural network detectors consider the background information to predict the identification. Applying segmentation to filter out the background force the deep network to focus on finding the features of the inner textures of the objects.

- Finding distinctive features: While characteristic features are essential for detection tasks, distinctive features are essential for identifying the objects, especially the objects with less visual information. In color images, texture features of a target object may be unique. With a single color channel, the outer shapes of the objects contribute more to the identification than the texture features. Attention mechanism is the key to guiding the deep network in focusing the distinctive characteristic features of the target object.

- Polymorphism problems: This study explores the relationship of various appearances of the target categories. There may be similar features shared among different categories. On the other hand, there are also different features among objects in the same category. Building a hierarchical structure for clustering the different appearances clears the ambiguity in each category.

Characteristic features is essential for detecting the target objects. The Segmentation-driven mechanism is proposed to guide the detector to focus on the characteristic features of the target objects. Chapter 3 performs the segmentation-driven idea on the protozoa domain to find the texture features. The main contributions of Chapter 3 to this problem are:

- Establishing Segmentation-driven RetinaNet, which filters out the background and detects the target objects.

- Introducing several data augmentation techniques to overcome the small number of data problem.

Even though there are at most five samples per life-cycle stage for training, we successfully train a practical system to detect, segment, and identify protozoa with high accuracy. Experimental results on the protozoa dataset show the effectiveness of the proposed segmentation-driven mechanism.

On the other hand, the distinctive features is essential for identifying the target objects. The Attention-driven mechanism is proposed to guide the network to focus on the distinctive features which also characterize the target objects. Chapter 4 performs the idea on the protozoa and Genome Profiling image domain to find the distinctive features. The main contributions of Chapter 4 are:

- Designing an prerequisite image enhancement process to enhance the connectivity of objects in grayscale images

- Sharper the segmentation results by stacking attention layers

- Establishing the Attention-driven RetinaNet that focuses on the distinctive features to improve the detection performance

Even though the case where there are few samples for training, the proposed method is still able to detect the spiddos and achieves mAP values of 0.29 and 0.54 on the two datasets of DNA Profiling images.

A hierarchical classifier is applied for the detection network to explore the similar and different features among the target objects. Chapter 5 introduces Segmentation-driven Hierarchical RetinaNet, which integrate the hierarchical classifier to the detection network for the various appearance problem of the object with less visual information. The proposed network is evaluated on the protozoa dataset. The proposed method captures the hierarchical relationship between the life-cycle stages and the species to improve the

detection performances. Experimental results show that our proposed network can learn the characteristic features of the target objects. There are five samples per life-cycle stage to train the detection methods. The proposed method achieves the highest mAP, precision, and recall values over the related works.

6.2 Future Work

This section gives some possible further directions of this study, which are suggested by the experiments and the recent progress of the computer vision.

Instance-level Segmentation

This dissertation performs semantic segmentation that clusters pixels into foreground or background. The semantic segmentation is efficient to eliminate the background, therefore boosting up the detection performance. However, the target object instances in the images should be treated independently. Instance-level segmentation, which clusters pixels into the specific object instance in details, is expected to guide the neural network detector to separate the features among instances appeared in the image. Instance-level segmentation is also expected to help the detector in case of overlapping objects.

Subcategory Data augmentation

A solution to overcome the lack of data problem is the data augmentation technique. In this dissertation, color transfer and rotation augmentation techniques are applied to train the neural network detectors. A more semantic augmentation is to generate more data sample with respect to the subcategories corresponding to the various appearances. Generative Adversarial Network (GAN) [103] or Variational Auto Encoder (VAE) [104] can be applied to generate more data samples.

Bibliography

[1] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 1, pp. 142–158, 2016.

[2] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[3] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2999–3007.

[4] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *IEEE International Conference on Computer Vision*, 2017, pp. 2980–2988.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE conference on computer vision and pattern recognition*, 2009, pp. 248–255.

[6] "An overview of deep learning in medical imaging focusing on mri," *Zeitschrift für Medizinische Physik*, vol. 29, no. 2, pp. 102–127, 2019.

[7] L. Brigato and L. Iocchi, "A close look at deep learning with small data," *CoRR*, vol. abs/2003.12843, 2020. [Online]. Available: https://arxiv.org/abs/2003.12843

[8] E. Togootogtokh and A. Amartuvshin, "Deep learning approach for very similar

objects recognition application on chihuahua and muffin problem," *CoRR*, vol. abs/1801.09573, 2018. [Online]. Available: http://arxiv.org/abs/1801.09573

[9] D. Paschalidou, L. V. Gool, and A. Geiger, "Learning unsupervised hierarchical part decomposition of 3d objects from a single RGB image," *CoRR*, vol. abs/2004.01176, 2020. [Online]. Available: https://arxiv.org/abs/2004.01176

[10] C. T. N. Suzuki, J. F. Gomes, and A. X. Falcao, "Automatic segmentation and classification of human intestinal parasites from microscopy images," *Biomedical Engineering*, vol. 60, pp. 803–812, 2013.

[11] "A robust technique based on invariant moments – anfis for recognition of human parasite eggs in microscopic images," *Expert Systems with Applications*, vol. 35, no. 3, pp. 728 – 738, 2008.

[12] F. Thung and I. S. Suwardi, "Blood parasite identification using feature based recognition," in *International Conference on Electrical Engineering and Informatics*, 2011, pp. 1–4.

[13] J. L. V. Noguera, H. L. Ayala, C. E. Schaerer, and M. Rolón, "Mathematical morphology for counting trypanosoma cruzi amastigotes," *XXXIX Latin American Computing Conference (CLEI)*, pp. 1–12, Oct 2013.

[14] F. B. Dazzo and B. C. Niccum, "Use of cmeias image analysis software to accurately compute attributes of cell size, morphology, spatial aggregation and color segmentation that signify in situ ecophysiological adaptations in microbial biofilm communities," *Computation*, vol. 3, no. 1, pp. 72–98, 2015.

[15] Z. Ji, K. J. Card, and F. B. Dazzo, "Cmeias jfrad: A digital computing tool to discriminate the fractal geometry of landscape architectures and spatial patterns of individual cells in microbial biofilms," *Microbial Ecology*, vol. 69, no. 3, pp. 710–720, 2015.

[16] C. Li, *Content-based Microscopic Image Analysis*, May 2016.

[17] P. J. P., F. A. X., and S. C. T. N., "Supervised pattern classification based on optimum-path forest," *International Journal of Imaging Systems and Technology*, vol. 19, no. 2, pp. 120–131, 2009.

[18] Y. S. Yang, D. K. Park, H. C. Kim, M.-H. Choi, and J.-Y. Chai, "Automatic identification of human helminth eggs on microscopic fecal specimens using digital image processing and an artificial neural network," *IEEE Transactions on Biomedical Engineering*, vol. 48, no. 6, pp. 718–730, 2001.

[19] R. Flores-Quispe, R. E. Patiño Escarcina, Y. Velazco-Paredes, and C. A. Beltrán Castañón, "Classification of human parasite eggs based on enhanced multitexton histogram," in *2014 IEEE Colombian Conference on Communications and Computing (COLCOM)*, 2014, pp. 1–6.

[20] Y. Zou, C. Li, K. Shirahama, T. Jiang, and M. Grzegorzek, "Environmental microorganism image retrieval using multiple colour channels fusion and particle swarm optimisation," in *IEEE International Conference on Image Processing (ICIP)*, 2016, pp. 2475–2479.

[21] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, Nov 2004. [Online]. Available: https://doi.org/10.1023/B:VISI.0000029664.99615.94

[22] S. Kosov, K. Shirahama, C. Li, and M. Grzegorzek, "Environmental microorganism classification using conditional random fields and deep convolutional neural networks," *Pattern Recognition*, vol. 77, pp. 248–261, 2018.

[23] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *CoRR*, vol. abs/1606.00915, 2016. [Online]. Available: http://arxiv.org/abs/1606.00915

[24] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, in *Proceedings of the Eighteenth International Conference on Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, pp. 282–289.

[25] K. Nishigaki, N. Amano, and T. Takasawa, "Dna profiling. an approach of systemic characterization, classification, and comparison of genomic dnas," *Chemistry Letters*, vol. 20, no. 7, pp. 1097–1100, 1991.

[26] A. Hingorani, V. Kuan, C. Finan, F. Kruger, A. Gaulton, S. Chopade, R. Sofat, R. MacAllister, J. Overington, H. Hemingway, S. Denaxas, D. Prieto, and J. P. Casas, "Flipping the odds of drug development success through human genomics," 2017. [Online]. Available: https://doi.org/10.1101/170142

[27] T. Kinebuchi, N. Idota, H. Tsuboi, M. Takaso, R. Bando, and H. Ikegaya, "The genome profiling method can be applied for species identification of biological materials collected at crime scenes," *BMC genetics*, vol. 20, no. 1, pp. 1–7, 2019.

[28] H. Nakamura, T. Muro, S. Imamura, and I. Yuasa, "Forensic species identification based on size variation of mitochondrial dna hypervariable regions," *International journal of legal medicine*, vol. 123, no. 2, pp. 177–184, 2009.

[29] K. Hamano, S. Ueno-Tsuji, R. Tanaka, M. Suzuki, K. Nishimura, and K. Nishigaki, "Genome profiling (gp) as an effective tool for monitoring culture collections: a case study with trichosporon," *Journal of microbiological methods*, vol. 89, no. 2, pp. 119–128, 2012.

[30] T. Shiraishi, K. Sekiguchi, and T. Ohmori, "Validation studies of 'oc-hemocatch' for the forensic identification of human blood," *Japanese Journal of Forensic Science and Technology*, vol. 7, pp. 159–165, 2003.

[31] H. Katagiri, C. Nishida, K. Terao, T. Yoshii, K. Matsumura, and Y. Uchimura, "Comparative studies of commercial fecal occult blood test kits for the identifica-

tion of human bloodstains," *Japanese Journal of Forensic Science and Technology*, vol. 14, no. 1, pp. 29–34, 2009.

[32] H. Tsutsumi, H. H. Htay, K. Sato, and Y. Katsumata, "Antigenic properties of human and animal bloodstains studied by enzyme-linked immunosorbent assay (elisa) using various antisera against specific plasma proteins," *Zeitschrift für Rechtsmedizin*, vol. 99, pp. 191–196, 1987.

[33] "Species identification of blood and bloodstains by enzyme-linked immunosorbent assay (elisa) using anti-human immunoglobulin kappa light chain monoclonal antibody," *Forensic Science International*, vol. 40, no. 1, pp. 85–95, 1989.

[34] J. E. Lygo, P. E. Johnson, D. J. Holdaway, S. Woodroffe, C. P. Kimpton, P. Gill, J. P. Whitaker, and T. M. Clayton, "The validation of short tandem repeat (str) loci for use in forensic casework," *International Journal of Legal Medicine*, vol. 107, pp. 77–89, 1994.

[35] T. Ono, S. Miyaishi, Y. Yamamoto, K. Yoshitome, T. Ishikawa, and H. Ishizu, "Human identification from forensic materials by amplification of a human-specific sequence in the myoglobin gene," *Acta Med Okayama*, vol. 55, no. 3, pp. 175–184, Jun 2001.

[36] H. Wittig, C. Augustin, A. Baasner, U. Bulnheim, N. Dimo-Simonin, J. Edelmann, S. Hering, S. Jung, S. Lutz, M. Michael, W. Parson, M. Poetsch, P. M. Schneider, G. Weichhold, and D. Krause, "Mitochondrial DNA in the Central European population. Human identification with the help of the forensic mt-DNA D-loop-base database," *Forensic Sci Int*, vol. 113, no. 1-3, pp. 113–118, Sep 2000.

[37] H. Nakamura, T. Muro, S. Imamura, and I. Yuasa, "Forensic species identification based on size variation of mitochondrial DNA hypervariable regions," *Int J Legal Med*, vol. 123, no. 2, pp. 177–184, Mar 2009.

[38] W. Parson, K. Pegoraro, H. Niederstätter, M. Föger, and M. Steinlechner, "Species identification by means of the cytochrome b gene," *Int J Legal Med*, vol. 114, no. 1-2, pp. 23–28, 2000.

[39] K. Imaizumi, T. Akutsu, S. Miyasaka, and M. Yoshino, "Development of species identification tests targeting the 16S ribosomal RNA coding region in mitochondrial DNA," *Int J Legal Med*, vol. 121, no. 3, pp. 184–191, May 2007.

[40] E. Naito, K. Dewa, H. Ymanouchi, and R. Kominami, "Ribosomal ribonucleic acid (rRNA) gene typing for species identification," *J Forensic Sci*, vol. 37, no. 2, pp. 396–403, Mar 1992.

[41] R. Makino, K. Kaneko, T. Kurahashi, T. Matsumura, and K. Mitamura, "Detection of mutation of the p53 gene with high sensitivity by fluorescence-based PCR-SSCP analysis using low-pH buffer and an automated DNA sequencer in a large number of DNA samples," *Mutat Res*, vol. 452, no. 1, pp. 83–90, Jul 2000.

[42] "Species-identification dots: a potent tool for developing genome microbiology," *Gene*, vol. 261, no. 2, pp. 243–250, 2000.

[43] C. Steger, "An unbiased detector of curvilinear structures," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 20, no. 2, pp. 113–125, 1998.

[44] ——, "Removing the bias from line detection," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1997, pp. 116–122.

[45] ——, "Extraction of curved lines from images," in *Proceedings of 13th International Conference on Pattern Recognition*, vol. 2, 1996, pp. 251–255.

[46] N. Suwa, H. Ikegaya, T. Takasaka, K. Nishigaki, and K. Sakurada, "Human blood identification using the genome profiling method," *Legal Medicine*, vol. 14, no. 3, pp. 121–125, 2012.

[47] R. Hirata, T. Takasaka, D. Miyamori, S. Ahmed, K. Sakurada, K. Nishigaki, and H. Ikegaya, "Use of the genome profiling method for the identification of saliva and sweat samples," *Japanese Journal of Forensic Science and Technology*, vol. 18, no. 1, pp. 79–83, 2013.

[48] T. Takasaka, K. Sakurada, T. Akutsu, K. Nishigaki, and H. Ikegaya, "Trials of the detection of semen and vaginal fluid rna using the genome profiling method," *Legal Medicine*, vol. 13, no. 5, pp. 265–267, 2011.

[49] M. Kouduka, D. Sato, M. Komori, M. Kikuchi, K. Miyamoto, A. Kosaku, M. Naimuddin, A. Matsuoka, and K. Nishigaki, "A solution for universal classification of species based on genomic dna," *International journal of plant genomics*, vol. (2007), pp. 1–8, 2007.

[50] D. Diwan, S. Komazaki, M. Suzuki, N. Nemoto, T. Aita, A. Satake, and K. Nishigaki, "Systematic genome sequence differences among leaf cells within individual trees," *BMC genomics*, vol. 15, no. 1, p. 142, 2014.

[51] D. Diwan, Y. Masubuchi, T. Furukawa, and K. Nishigaki, "Ordered genome change of plant and animal body cells revealed by the genome profiling method," *FEBS letters*, vol. 590, no. 14, pp. 2119–2126, 2016.

[52] S.-H. Yang, J.-K. Cho, S.-Y. Lee, O. D. Abanto, S.-K. Kim, C. Ghosh, J.-S. Lim, and S.-G. Hwang, "Isolation and characterization of novel denitrifying bacterium geobacillus sp. sg-01 strain from wood chips composted with swine manure," *Asian-Australasian journal of animal sciences*, vol. 26, no. 11, p. 1651, 2013.

[53] M. Naimuddin, T. Kurazono, and K. Nishigaki, "Commonly conserved genetic fragments revealed by genome profiling can serve as tracers of evolution," *Nucleic acids research*, vol. 30, no. 10, p. e42, 2002.

[54] K. Enomoto, Y. Hatakeyama, N. Nishimoto, S. Miyake, H. Oda, M. Takahashi, Y. Yamamoto, and H. Iwano, "Analysis of taxonomic inference of 40 bacillus

thuringiensis serotypes using genome profiling," *Journal of Insect Biotechnology and Sericology*, vol. 84, no. 1, pp. 1–7, 2015.

[55] T. Fujino, H. Wityi, T. Nomoto, K. Nishigaki, T. Kondo, A. Limsakul, and S. Davison, "Application of genome profiling method to the study of closely related species of stenopsyche in japan, viet nam and thailand," *Biology Inland Water Supplement 2*, vol. 2, pp. 19–26, 2012.

[56] S. Ahmed, M. Komori, S. Tsuji-Ueno, M. Suzuki, A. Kosaku, K. Miyamoto, and K. Nishigaki, "Genome profiling (gp) method based classification of insects: congruence with that of classical phenotype-based one," *PloS one*, vol. 6, no. 8, p. e23963, 2011.

[57] P. Thanakiatkrai and T. Kitpipit, "Meat species identification by two direct-triplex real-time pcr assays using low resolution melting," *Food chemistry*, vol. 233, pp. 144–150, 2017.

[58] S. A. K. Nishigaki, "Error-robust nature of genome profiling applied for clustering of species demonstrated by computer simulation," *International Journal of Bioengineering and Life Sciences*, vol. 1, no. 5, pp. 1–7, May 2007.

[59] H. Sharma, F. Ohtani, P. Kumari, D. Diwan, N. Ohara, T. Kobayashi, M. Suzuki, N. Nemoto, Y. Matsushima, and K. Nishigaki, "Familial clustering of mice consistent to known pedigrees enabled by the genome profiling (gp) method," *Biophysics*, vol. 10, pp. 55–62, 2014.

[60] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.

[61] R. Vaillant, C. Monrocq, and Y. Le Cun, "Original approach for the localisation of objects in images," *IEEE Proceedings-Vision, Image and Signal Processing*, vol. 141, no. 4, pp. 245–250, 1994.

[62] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2001, pp. 511–518.

[63] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *PAMI*, vol. 32, no. 9, pp. 1627–1645, 2010.

[64] P. F. Felzenszwalb, R. B. Girshick, and D. McAllester, "Cascade object detection with deformable part models," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010, pp. 2241–2248.

[65] J. Yan, Z. Lei, L. Wen, and S. Z. Li, "The fastest deformable part model for object detection," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2497–2504.

[66] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[67] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5987–5995.

[68] G. Ghiasi, T. Lin, R. Pang, and Q. V. Le, "NAS-FPN: learning scalable feature pyramid architecture for object detection," *CoRR*, vol. abs/1904.07392, 2019. [Online]. Available: http://arxiv.org/abs/1904.07392

[69] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," 2017, cite arxiv:1704.04861. [Online]. Available: http://arxiv.org/abs/1704.04861

[70] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6848–6856.

[71] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer.

[72] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *International journal of computer vision*, vol. 104, no. 2, pp. 154–171, 2013.

[73] K. E. Van de Sande, J. R. Uijlings, T. Gevers, and A. W. Smeulders, "Segmentation as selective search for object recognition," in *IEEE International Conference on Computer Vision*, 2011, pp. 1879–1886.

[74] T.-Y. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie, "Feature pyramid networks for object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 936–944.

[75] A. Shrivastava, H. Mulam, and R. Girshick, "Training region-based object detectors with online hard example mining," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 761–769.

[76] A. Shrivastava, R. Sukthankar, J. Malik, and A. Gupta, "Beyond skip connections: Top-down modulation for object detection," *arXiv preprint arXiv:1612.06851*, 2016.

[77] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," *arXiv preprint arXiv:1312.6229*, 2013.

[78] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[79] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, "Dssd: Deconvolutional single shot detector," *arXiv preprint arXiv:1701.06659*, 2017.

[80] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 779–788.

[81] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7263–7271.

[82] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: common objects in context," *CoRR*, vol. abs/1405.0312, 2014. [Online]. Available: http://arxiv.org/abs/1405.0312

[83] P. O. Pinheiro, T.-Y. Lin, R. Collobert, and P. Dollár, "Learning to refine object segments," in *European Conference on Computer Vision*, 2016, pp. 75–91.

[84] C. Li, K. Shirahama, and M. Grzegorzek, "Application of content-based image analysis to environmental microorganism classification," *Biocybernetics and Biomedical Engineering*, vol. 35, no. 1, pp. 10 – 21, 2015.

[85] B. Y. Yu, C. Elbuken, C. L. Ren, and J. P. Huissoon, "Image processing and classification algorithm for yeast cell morphology in a microfluidic chip," *Journal of Biomedical Optics*, vol. 16, no. 6, pp. 66 008–66 018, 2011.

[86] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *IEEE International Conference on Computer Vision*, 2017, pp. 618–626.

[87] E. Reinhard, M. Adhikhmin, B. Gooch, and P. Shirley, "Color transfer between images," *IEEE Computer graphics and applications*, vol. 21, no. 5, pp. 34–41, 2001.

[88] M. Everingham and J. Winn, "The pascal visual object classes challenge 2007 development kit," *University of Leeds, Tech. Rep*, 2007.

[89] Y. Ji, H. Zhang, and Q. Jonathan Wu, "Salient object detection via multi-scale attention cnn," *Neurocomputing*, vol. 322, pp. 130–140, 2018.

[90] z. huang, W. Ke, and D. Huang, "Improving object detection with inverted attention," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, March 2020.

[91] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186.

[92] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," vol. abs/2010.11929, 2020. [Online]. Available: https://arxiv.org/abs/2010.11929

[93] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.

[94] K. Pho, M. K. M. Amin, and A. Yoshitaka, "Segmentation-driven retinanet for protozoa detection," in *IEEE International Symposium on Multimedia*. IEEE, 2018, pp. 279–286.

[95] J. Levatić, D. Kocev, and S. Džeroski, "The importance of the label hierarchy in hierarchical multi-label classification," *Journal of Intelligent Information Systems*, vol. 45, pp. 247–271, Oct 2015.

[96] G. Obozinski, G. Lanckriet, C. Grant, M. I. Jordan, and W. S. Noble, "Consistent

probabilistic outputs for protein function prediction," *Genome Biology*, vol. 9, no. 1, p. S6, 2008.

[97] G. Valentini, "True path rule hierarchical ensembles for genome-wide gene function prediction," *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, vol. 8, no. 3, pp. 832–847, 2011.

[98] C. Vens, J. Struyf, L. Schietgat, S. Džeroski, and H. Blockeel, "Decision trees for hierarchical multi-label classification," *Machine learning*, vol. 73, no. 2, p. 185, 2008.

[99] H. Blockeel, L. Schietgat, J. Struyf, S. Džeroski, and A. Clare, "Decision trees for hierarchical multilabel classification: A case study in functional genomics," in *European Conference on Principles of Data Mining and Knowledge Discovery*. Springer, 2006, pp. 18–29.

[100] R. Cerri, R. C. Barros, and A. C. De Carvalho, "Hierarchical multi-label classification using local neural networks," *Journal of Computer and System Sciences*, vol. 80, no. 1, pp. 39–56, 2014.

[101] X. Fan, "Efficient multiclass object detection by a hierarchy of classifiers," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2005, pp. 716–723.

[102] M. B. Bueno, X. Giró-i Nieto, F. Marqués, and J. Torres, "Hierarchical object detection with deep reinforcement learning," *Deep Learning for Image Processing Applications*, vol. 31, no. 164, p. 3, 2017.

[103] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, ser. NIPS'14. Cambridge, MA, USA: MIT Press, 2014, p. 2672–2680.

[104] D. P. Kingma and M. Welling, "An introduction to variational autoencoders,"

CoRR, vol. abs/1906.02691, 2019. [Online]. Available: http://arxiv.org/abs/1906.02691

Publications

Journals

[1] Khoa Pho, Misato Baba, Yutaro Narukawa, Yuto Tsutsumi, Kiyoshi Yasukawa, and Atsuo Yoshitaka: Species Identification Dots Detection in TGGE Images toward Automatic Genome Profiling, *IEEE Access*, 2022. (**Under Review**)

[2] Khoa Pho, Muhamad Kamal Mohammed Amin, and Atsuo Yoshitaka: Segmentation-driven Hierarchical RetinaNet for Detecting Protozoa in Micrograph, Vol. 13, No. 03, pp. 393-413, Special issue, *International Journal of Semantic Computing*, 2019.

International Conferences

[3] Khoa Pho, Muhamad Kamal Mohammed Amin, and Atsuo Yoshitaka: Segmentationdriven RetinaNet for Protozoa Detection, pp. 279-286, *IEEE International Symposium on Multimedia (ISM)*, 2018.

[4] Khoa Pho, Takafumi Hirase, Muhamad Kamal Mohammed Amin, and Atsuo Yoshitaka: Protozoa Identification using 3D Geometric Multiple Color Channel Local Feature, pp. 37-42, *IEEE International Conference on Knowledge and Systems Engineering (KSE)*, 2018.

Domestic Conferences

[5] Khoa Pho, Misato Baba, Yutaro Narukawa, Yuto Tsutsumi, Kiyoshi Yasukawa, and Atsuo Yoshitaka: An Image Enhancement Method for TGGE Images for Genome Profiling, *Joint conference of Hokuriku chapters of Electrical Society (JHES)*, 2021.

[6] Khoa Pho, Takafumi Hirase, Muhamad Kamal Mohammed Amin, and Atsuo Yoshitaka: Protozoa Identification using 3D Geometric Multiple Color Channel Local Feature, *Joint conference of Hokuriku chapters of Electrical Society (JHES)*, 2018.