2021

Deep Learning Based Microatoll Detection From Drone Images

Zhexuan Zhou

Unless otherwise indicated, the views expressed in this thesis are those of the author and do not necessarily represent the views of the University of Wollongong.

Deep Learning Based Microatoll Detection From Drone Images

Zhexuan Zhou

This thesis is presented as part of the requirements for the conferral of the degree:

Master of Philosophy

Supervisor:
A/Prof. Lei Wang

Co-supervisor:
Dr. Luping Zhou

The University of Wollongong
School of Computing and Information Technology

October 10, 2021

This research has been conducted with the support of an Australian Government Research Training Program Scholarship.

Declaration

I, *Zhexuan Zhou*, declare that this thesis is submitted in partial fulfilment of the requirements for the conferral of the degree *Master of Philosophy*, from the University of Wollongong, is wholly my own work unless otherwise referenced or acknowledged. This document has not been submitted for qualifications at any other academic institution.

Zhexuan Zhou

October 10, 2021

Abstract

Visual object detection has made significant progress with the advent of deep neural networks and has been extensively applied. This thesis reports a novel application that aims to detect individual microatolls, which are circular coral colonies, from island images captured by drones. We first describe data collection and labelling to create a novel microatoll dataset for the microatoll detection task from drone images. Upon this dataset, the state-of-the-art object detectors are then evaluated for this task. To better integrate a detector with the characteristic of microatolls, we propose a modified detector called Microatoll-Net. It actively extracts features from the surrounding area of a microatoll to differentiate it from distractors to improve detection. Multiple ways to incorporate this information into the detector are designed. The experimental study shows the efficacy of the proposed Microatoll-Net, especially on the most challenging area for detection. Besides, in geographical research, the position of a microatoll is more important than its size. It means that we shall pay more attention to detecting the centre of a microatoll instead of generating its bounding box. Motivated by this, we transform this object detection task into an object centre detection task.

Acknowledgments

I would like to express my sincere gratitude to my supervisor, Associate Prof. Lei Wang, to support my M.Phil study and related research for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my M.Phil study. I also would like to thank my co-supervisor, Dr Luping Zhou, for her constant support and constructive suggestions.

Besides my supervisors, I would like to acknowledge Associate Prof. Sarah M. Hamylton of the School of Earth, Atmospheric and Life Sciences, University of Wollongong. She offered great help in the labelling procedure, enhancing dataset reliability and evaluating performance. Without her help, the difficulty of labelling data will be unimaginable.

Special thanks to my fellows in my group and lab. Thanks for your support in the past two years. In particular, I would like to thank Biting Yu, Yu Ding, Zhongyan Zhang and Saimnunur Rahaman for giving me hints in completing experiments and comments on the annual research review rehearsals. I am also grateful to Jiayin Lin and Bela Chakraborty for help me out with polishing my thesis. I also thank my friends Jiayin Lin, Yunshu Zhu and Ting Song for every moment we have shared and every footprint we have walked. All of you enlighten my M.Phil study and make it colourful.

Last but not least, I must express my very profound gratitude to my parents Jianfeng Zhou and Qiu'e Huang and my girlfriend Peng Zhou to provide me with continuous support and encouragement throughout the years of my M.Phil course. This accomplishment would not have been possible without them. Thank you.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Contents

1.1 Research Background

Unmanned Aerial Vehicle (UAV), also known as a drone, is an aircraft without a human pilot. The UAV is a part of an unmanned aircraft system, including a ground-based controller and a system between them for their communication (an illustration is shown in Fig. 1.1). Such a structure is very convenient for users to operate, so people with a drone licence can satisfactorily complete the route design, take-off and recovery. Therefore, it is advantageous to use a drone to collect data as high mobility allows it to collect data more efficiently. Also, it can be used to collect data in places difficult for humans to reach. Specifically, the planability of the course of action makes the collected data more meaningful and accurate. The high-resolution camera mounted on a drone is able to capture and retain sufficient details of the target area. The mobility, plannability, and capturability make a drone a cost-effective way of gathering information compared with a fieldwork in persons.

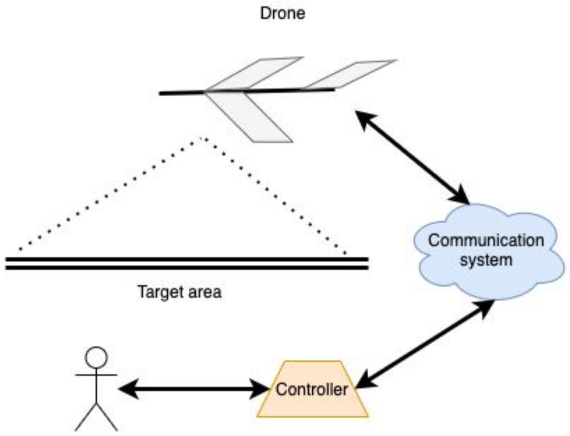


Figure 1.1: This diagram illustrates an unmanned aircraft system. A human operator operates the controller on the ground, and the signal is sent to the communication system, which could be a part of the controller or an online system. After that, the channel between the drone and the communication system exchanges user information and data such as field images.

The widespread usage of drones also covers geological research, where environmental scientists conduct field research with a drone to assemble photographs. As aforementioned, it is advantages to use drones for geological research. Accordingly, it is quite intuitive to apply it to the study of sea-level change through microatolls. Microatolls are boulder-like (an example is shown in Fig. 1.2) corals of the genus Porites that act as natural recorders of sea-level [1]. As described in the work of Scoffin et al. [2], microatolls record the tidal changes by the biological behaviours: as it is under the water, a microatoll grows in horizontal and vertical directions and forms a hemisphere shape. Whereas its exposures to the air could stop its vertical growth, it continues to be broader, resulting in a circular shape. Besides, tides changes will affect their exterior because tides may flip microatolls upside down and bring sea-bed changes, and these uncertainties cause microatolls to have a variety of shapes, such as significant breakages across themselves and forming a classical horse-shoes shape. Over the period, the stable sea-level creates a suitable environment for microatolls to become broader to form microatoll pavements.

The microatoll pavements lose the ability to record sea-level change because of forming coral reef lagoon.

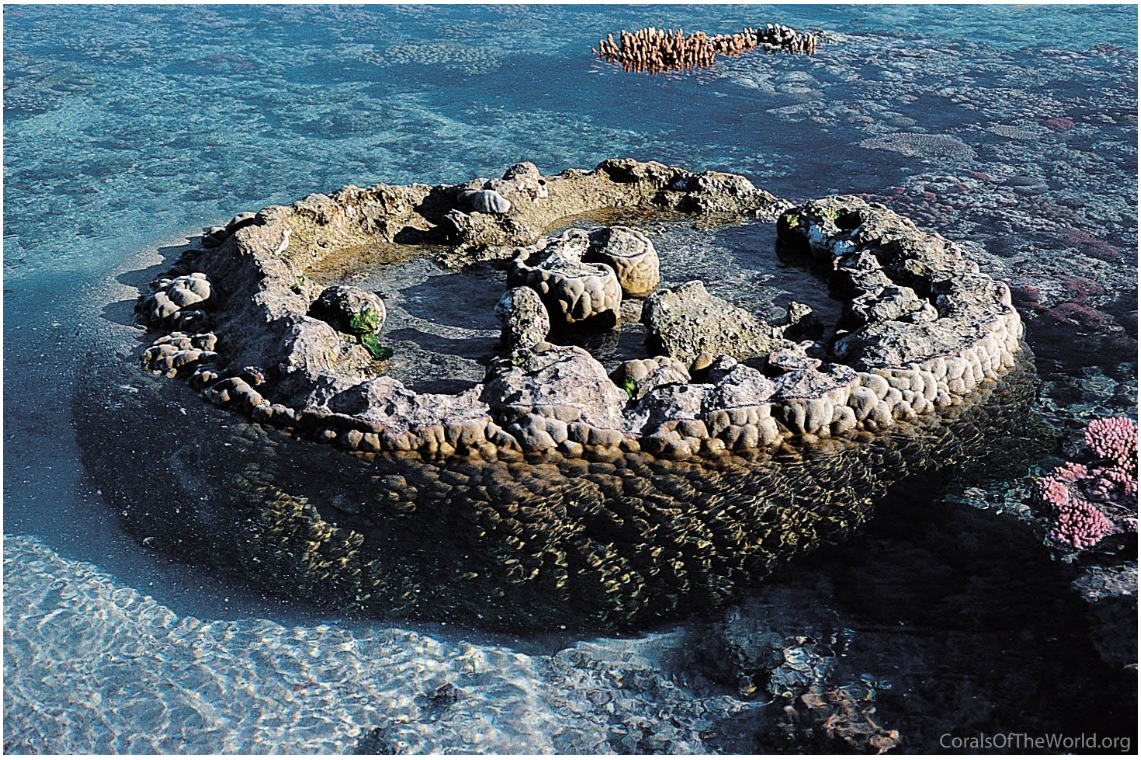


Figure 1.2: This figure shows the sample of a microatoll. Image courtesy of the organization of Coral of The World.

As a drone is efficient for high-quality image collection, the microatoll labelling naturally becomes a critical step to narrow the gap between raw fieldwork image data and real-world application. Technically, locating microatolls from given drone images by computers can be formulated as a computer vision task, and deep learning methods have shown their outstanding capability of handling various artificial intelligence related tasks. In 2016, Alpha-GO, the first computer program that defeated a professional human Go player, shows that robust and well designed artificial intelligence systems could be exploited to perform tasks that need human intelligence. Since then, artificial intelligence has attracted intensive attention. Tracing back to 2012, the AlexNet [3] performed considerably better than the previous state-of-the-art methods in object classification task, starting a deep learning based revolution in the field of computer vision via convolutional neural networks (CNN) [4]. It automatically extracts robust and generalised feature representation for image recognition tasks. Additionally, the feature representation from CNN also benefits other computer vision tasks.

Specifically, locating the microatolls from drone images could be categorised as an object detection task. Object detection is one of the vital tasks in computer vision. It maps an image to a set of bounding boxes indicating the location and category of objects of interest. A key factor of object detection is the quality of feature representation. Compared with those detectors that utilise hand-craft features (i.e. histogram of oriented gradients (HOG) [5] and scale-invariant feature transform (SIFT) [6]), the deep learning based object detection methods have notable advantages in speed, efficiency and accuracy. HOG and SIFT rely on human experts to build and extract hand-crafted features (also named shallow features, such as colours, edges and corner in images). In contrast, CNN could automatically extract deep features (such as pixel relationships) efficiently and abundantly. It significantly increases the quality of feature representation for object detection. Therefore, this thesis employs CNN to respond to the challenge of detecting microatoll from drone images.

1.2 Thesis Organisation

For detecting microatolls, this thesis is organised into five chapters. The first chapter is an introduction that provides the research background, aims and significance. The second chapter introduces the significant literature in the field, including the geographic research for microatoll dataset construction, object detection development in the past two decades, their suitable application scenarios and limitations. The third chapter describes the microatoll dataset construction, investigates modern state-of-the-art object detection methods on the constructed microatoll dataset and proposes Microatoll-Net to better incorporate the surrounding information of a microatoll for its detection. The fourth chapter presents a novel detector based on the generative adversarial network for microatoll detection from drone images. The motivation is that the microatoll centres can well indicate their locations related to their circular shape. This model removes the need for manually designed anchor settings (introduced in Chapter 2) in modern object detection methods.

Also, it does not need to use the non-maximum suppression to merge the bounding boxes after the object detection procedure. In the final chapter, the research outcomes of this thesis are summarised, and future research directions are discussed.

1.3 Research Aim

This thesis aims to address the challenge in the task of microatoll detection from drone images. As a challenging issue, detecting objects from drone images has recently attracted research attention [7]. Different to the benchmark image datasets [8, 9] which include common objects in context, objects in drone images are tiny, difficult to find, and could be confused with background and other objects. This characteristic prevents popular object detectors from simply being transplanted to detect objects viewed from drones, although they achieved great success in ordinary detection settings. The recent literature [7] has reported that the performances of the state-of-the-art object detectors became worse in drone image datasets than their performances in the common benchmark image datasets. Detecting microatolls from drone images shares the same challenges in drone-based generic image detection. This thesis aims to enhance the efficiency and accuracy of detection microatolls from drone images to help geographic research.

1.4 Research Significance

The study of climate has attracted more attention, especially in the area of its history. The history of climate change helps understand alterations in the environment, such as temperature, animal migrations, and sea-level changes. The sea-level changes can reflect climate changes because ice melting in the Antarctic and Arctic is powerful evidence of global temperature changes. However, the total amount of the ice is difficult to accurately estimate. Therefore, it is a simpler and more effective method to estimate the history of temperature fluctuation based on changes in sea-level.

In geographic research, there are many ways to investigate and study the history of sea-level changes, such as using carbon dating [10] to estimate the age of marine fossils to roughly determine the sea-level. Compared with other methods, it is more accurate and convenient to use microatolls to determine the sea-level history. Microatolls are relatively easy to locate compared to the identification methods that need fossils or rocks. They live in open waters, have obvious characteristics and do not require geological mining. Meanwhile, its growth is sensitive to sea-level changes and this enables its growth to reflect sea-level changes with minimum errors (within three centimeters).

Locating microatolls is straightforward but still tricky because of their habitats. They live in open water apart from the coastline, and the field research is inefficient and not convenient. By using drones in geological research, the high-resolution geographic photographs allow researchers to locate microatolls before field research. Obviously, manually detecting microatolls from drone images is not the best solution in terms of efficiency and accuracy.

This thesis is to employ deep learning methods to automatically detect microatolls from drone images. Applying this method to high-resolution drone images is a key step to improve the efficiency of geographic fieldwork. With the help of deep learning methods, the data collection and annotation procedures in microatoll based sea-level changes research will be more efficient and accurate.

1.5 Contributions

This thesis focuses on microatoll detection from drone images. There are two main contributions in this thesis.

First, a novel drone image based microatoll dataset is constructed for such a microatoll detection task. Meanwhile, the mainstream object detection frameworks (i.e. Faster R-CNN [11] and SSD [12]) are investigated to evaluate their detection performances and the challenges of transplanting them from common datasets to our microatoll dataset are highlighted. The experimental study upon the dataset shows that the surrounding information outside the truth microatolls significantly differs from that of the false positives.

Therefore, Microatoll-Net is proposed to incorporate the surrounding information in the training procedure. Consequently, the proposed Microatoll-Net surpasses the mainstream detector (Faster R-CNN) in the most challenging area. More than just seeking a better result, this thesis also investigates different variants of Microatoll-Net to find a suitable feature combination. Chapter 3 presents the microatoll dataset construction, investigates the state-of-the-art object detection methods, and proposes Microatoll-Net.

Second, Faster R-CNN and proposed Microatoll-Net rely on the anchors to detect microatolls. Meanwhile, the primary task of this thesis is to locate the position (instead of the shape) of microatolls from drone images for field research. Inspired by this observation, this thesis designs a model that focuses on the location of microatolls during detection. The proposed model is named Microatoll-GAN. It takes advantages of recent research on infrared small object segmentation [13] to reduce miss detection and false negative by two parallel networks. The proposed Microatoll-GAN outputs centre density maps for the input drone images to indicate microatoll locations. Meanwhile, although the model focuses on the location, the centre density can roughly illustrate the sizes of microatolls. More details of the model are presented in Chapter 4.

1.6 Publications

1. Sarah M. Hamylton, Zhexuan Zhou and Lei Wang (2020) What Can Artificial Intelligence Offer Coral Reef Managers? Frontiers in Marine Science 7:603829. doi: 10.3389/fmars.2020.603829

Chapter 2

Literature Review

Contents

Deep learning methodology is a data-driven approach. It heavily relies on massive data to learn the patterns from the dataset to achieve its aim. Hence, only reliable annotations could help the deep learning models learn the efficacious data patterns and evaluate their performances. However, unlike the existing common object datasets [8, 9] or drone image dataset [7], microatolls are rare objects and are not included in these datasets. Thus, building a specific dataset on microatolls is inevitable. Also, it is necessary to carry a complete view of object detection literature to build a strong baseline for microatoll detection. Consequently, this chapter is organised as follows. First, it presents the literature on existing datasets and microatolls for constructing a reliable microatoll dataset from given drone images. Second, the development of object detection methods from hand-crafted features based to deep learning based methods is reviewed. Finally, the importance of dataset construction and the reasons for baseline model selection are concluded.

2.1 Dataset Construction

2.1.1 Existing Datasets on Object Detection

The literature of existing object detection datasets reflects the importance and necessity of constructing a specific microatoll dataset for this thesis. In object detection challenges, many well-known datasets and benchmarks have been released in the past ten years, such as PASCAL VOC Challenges [8], MS-COCO Detection Challenge [9] and VisDrone-DET Challenge [7]. For convenience, name them VOC, COCO and VisDrone dataset in short, respectively. Many object detection algorithms have been developed based on them. Therefore, it becomes necessary to dig out their relationship with microatolls.

VOC challenge is one of the most important competitions in the development of computer vision. It includes multiple tasks, such as image classification, object detection, semantic segmentation and action detection. Two versions (VOC07 and VOC12) are most used in the object detection task. The former contains about 5,000 images with 12,000 annotations, and the latter has 11,000 images with 27,000 annotations in the latter. In terms of categories, it contains four classes: person, animal, vehicle and indoor, and each of the main classes includes several subclasses. The names of categories clearly indicate that microatolls are not included in the dataset. Compared with VOC, COCO is the most challenging object detection dataset available today. It contains 164,000 images with 897,000 annotations from 80 categories. Similarly, there are 12 super-categories: accessory, animal, appliance, electronic, food, furniture, indoor, kitchen, outdoor, person, sports and vehicle. Although the number of categories is more than of VOC, it can be confirmed that COCO does not consist of microatolls either. In terms of VisDrone, it is the most challenging object detection dataset for drone images. There are 8,599 images (6,471 for training, 548 for validation and 1,580 for testing) with 540,000 annotations from 10 categories: pedestrian, person, car, van, bus, truck, motor, bicycle, awning-tricycle, and tricycle. Again, microatolls are not included in this dataset.

In sum, the microatoll class is not in the commonly used object detection datasets. Constructing a specific microatoll dataset is necessary. The existing literature gives criteria to label and store image annotations. This thesis follows the VOC format that stores the top-left and button-right coordinators of the ground truth bounding boxes as annotations for the microatoll detection task.

2.1.2 Microatoll Morphological Study for Microatoll Dataset Construction

Microatolls are boulder-like corals of the genus Porites that act as natural recorders of sea-level [1]. Scoffin and Stoddart [2] described them as coral colonies with dead, flat tops and living perimeters. Meanwhile, they also highlighted how the growth of microatoll is affected by tidal changes. According to their fieldwork research, individual microatolls in drone images should have a circle shape. Also, their margins (living perimeters) are generally darker than the centres because the margins are alive. Taking Fig. 2.1 as an example, the forming process of a typical microatoll is described. For convenience, λ_i is used to indicate the sea-level at some time. Suppose a microatoll is in its habitat under sea-level λ_1 (a lower water-level). In the beginning, the microatoll grows up in both vertical and horizontal directions to form a hemispherical shape. Once it reaches the water-air surface λ_1, the vertical growth will stop while the horizontal growth remains the same. Therefore, the microatoll forms a platform that is roughly the same as the smaller red circle shown in the figure. After that, the horizontal growth continues. When the water-level increases to λ_2 (water-level higher than λ_1), the platform in λ_1 is a dead top that stays as before, but its alive margin (because the margin is still under λ_1) will continue in both vertical and horizontal directions. Thus, there is a circle higher than the dead top platform within two red circles. This example indicates that the periphery of live coral needs to remain submerged underwater for much of the tidal cycle, so living microatolls are not found on exposed areas of coral reef top platforms, such as inside islands. As shown in Fig. 2.2, the area outside the red circle may contain microatolls. Therefore, ensuring the specific areas that may contain microatolls will speed up the labelling process by removing those images that do not contain the habitat of microatolls.

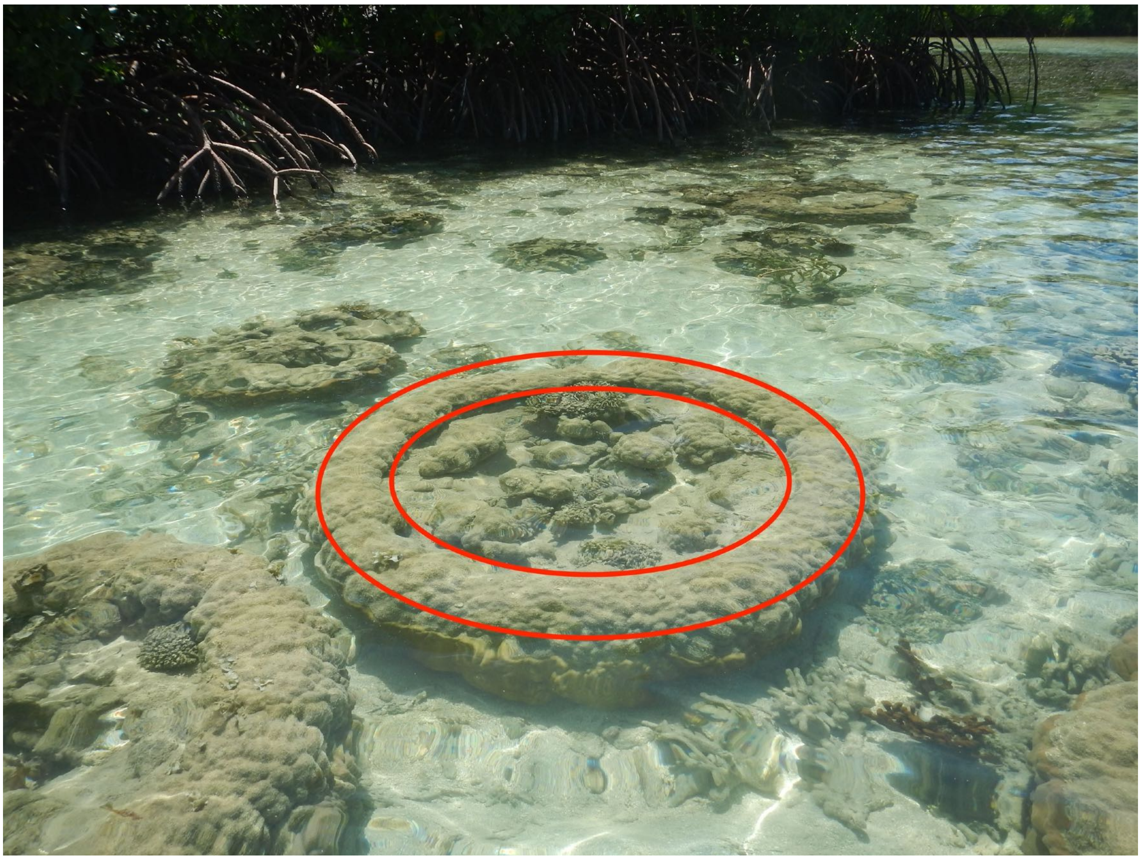


Figure 2.1: The microatoll example. As the image shows, the part within the two red circles is the living margin, which should be dark in the drone image. The smallest red circle is the dead top which has the lowest height. The irregular part of the dead top is due to the changes of sea-level.

Additionally, some individual microatolls may have breakages or form the shape of horseshoe due to the riverbed changes. These two shapes are typical in drone images, so figuring out their forming process would benefit the annotation process. According to the work of Scoffin et .al [2], the breakage of the microatoll skeleton into segments is because the sandy substrate is removed from the periphery of a large microatoll, then a disequilibrium will affect the continue growth trend. Therefore, those microatolls have decomposed into multiple small segments but still present the circular feature of an individual microatoll that can still be regarded as individual microatolls. Note that the horseshoe-shaped microatoll is a specific breakage. Because of the fluctuation of water-level changes and unstable riverbed, microatolls could have various shapes. Understanding the causes of cracks in microatolls will help reduce confusing labels when the annotating dataset.

Meanwhile, one type of microatolls named microatoll pavement is not the research target of this thesis. It is formed by individual microatolls that stay in a stable sea-level environment that reduces research value in sea-level change research.

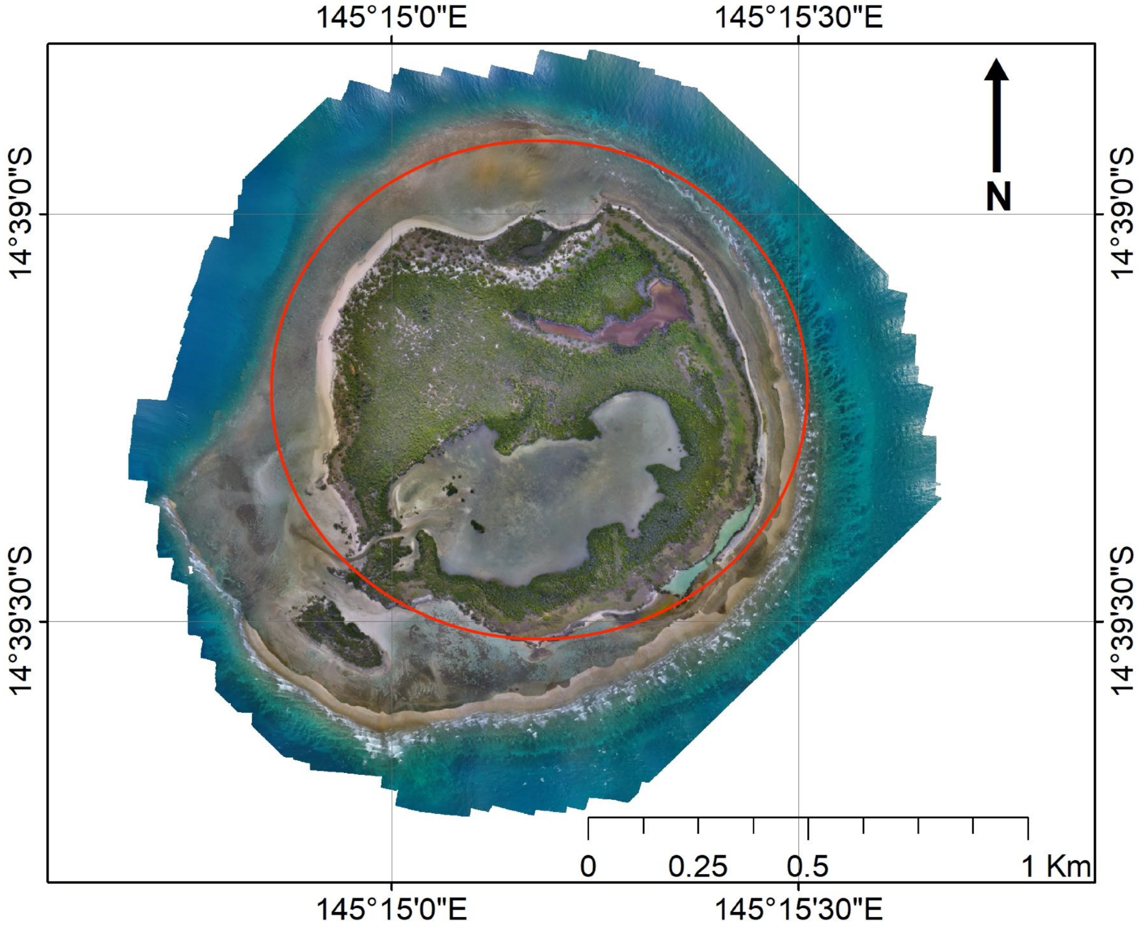


Figure 2.2: The areas of Nymph island that may contain microatolls. The areas outside the red circle have a great chance to contain microatolls because they have similar characteristics to microatoll habitats.

2.2 Object Detection

The year 2014 is a demarcation to split the traditional object detection and deep learning based detection, as shown in Fig. 2.3. Traditional object detection methods rely on handcrafted features generated and designed by human experts, while deep learning methods take advantages of convolutional neural network (CNN) to extract features automatically. Existing deep learning methods can be generally grouped into two- and one-stage ones. The former frames the detection as a coarse-to-fine procedure while the latter frames it as a completed in one network.

2.2.1 Hand-crafted Feature based Methods

Before AlexNet [3], hand-craft feature extraction methods were the most successful method in feature representation for many years. Those hand-craft features are robust, invariant to outliers such as lighting, scale changes, and background clutter, and have low computational costs. In sharp contrast to the advantages, the hand-craft feature is time-consuming, strongly relays on the experience of human experts, and they are not generalising to other tasks. Therefore, the hand-craft feature is also named shallow feature. Those popular methods in extracting shallow feature are Haar features [14], scale-invariant feature transform (SIFT) [3], a histogram of gradients (HOG) [5] and colour histograms. While in the early face detection task, Haar features are the most popular method used in Viola-Jones Detector [14], which is one of the most famous face detectors. It utilises sliding-windows to extract a large number of Haar features as input for a set of AdaBoost [15] classifiers. The cascade architecture connects weak detectors to be stronger detector and enhances the detection efficiency simultaneously. The Haar feature considers adjacent rectangular areas at specific positions in the detection windows, sums the pixel intensities in each area, and then calculates the difference between these sums. Finally, this difference is used to classify different parts of the image. Although Viola-Jones Detector is an “old school” detector, it highlights the three fundamental object detection problems: selection of detecting area, feature extraction, and classification, guiding the direction and procedure for later object detection methods.

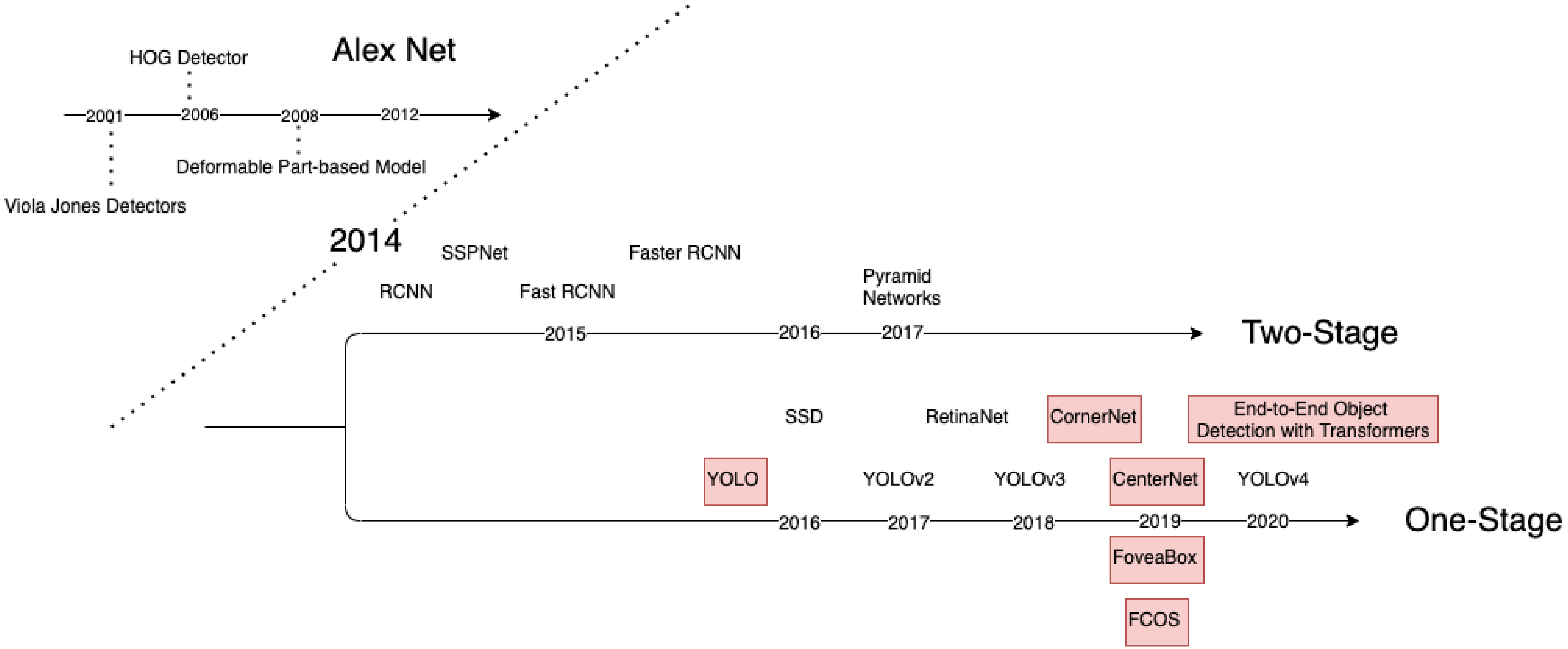


Figure 2.3: This figure presents the milestone of object detection during its development. It highlights the demarcation which separates the traditional and deep learning detection methodologies. The following methods are shown in this figure: Viola Jones Detectors [14], HOG Detector [5], Deformable Part-based Model [17], AlexNet [3], R-CNN [18], SSPNet [19], Fast R-CNN [20], Faster R-CNN [11], Pyramid Networks [21], YOLO (v1 to v4) [22–25], SSD [12], RetinaNet [26], CornerNet [27], CenterNet [28, 29], FoveaBox [30], FCOS [31] and End-to-End Object Detection with Transformers [32]. Those colored are also the Anchor Free methods.

2.2.2 Deep Learning based Object Detection

Critical layers in convolutional neural network

In order to have a better view of the theory of CNN, this section starts from the key layers of CNN. As mentioned in Section 2.2.1, the hand-crafted features have disadvantages of being time-consuming and not generalizing from one task to the others. Furthermore, they require experience from human experts. CNN well addresses these problems. In 1998, LeCun et al. [4] proposed LeNet to classify hand-written digits on MNIST [16] dataset. Although its structure (2 convolutional layers, 2 downsampling layers and 3 fully connected layers) is simple, it is introduced as the pioneer of deep learning methods.

As seen in LeNet, a CNN architecture is formed by a stack of distinct layers to transform input volumes to an output volume according to different purposes [4]. The convolutional layer is the core building block. It has a set of learnable filters with a small receptive field but extended to entire input volumes. The receptive field indicates the area where a filter learns from the input, and each of the filters presents the feature that they learn from input volumes. The size of output from CNN is controlled by output depth (the number of neurons or filters in the layer), stride (of filters going through entire input volumes) and padding (output size control). Another important layer is the pooling layer, which performs nonlinear sampling. As the most commonly used pooling method, Max-pooling partitions the input image into a set of non-overlapping rectangles and, for each sub-region, outputs the maximum value. At the same time, the Average-pooling acts the same but outputs the average value. An activation layer usually follows the convolutional layer and pooling layer to add non-linearity to help the neural network learn complex data patterns. Taking ReLU [33] as an example, it actively removes negative values from feature maps by setting them to zero: $f(x) = max(0,x)$. It increases the nonlinear properties of the decision function and the overall network without affecting the receptive fields of the convolution layer. Fully connected layers are the high-level reasoning in the neural network computed as an affine transformation, with matrix multiplication followed by a bias offset. The loss function indicates how to train the neural network, so it is usually the last neural network layer. Various loss functions appropriate for different tasks may be used, such as cross-entropy loss for multi-class prediction and Focal loss [26] to solve the foreground and background imbalance problem. These layers lay the foundation for a powerful convolutional neural network backbone such as ResNet [34], VGG [35] and DarkNet [36].

Those famous backbones make significant contributions to feature extraction. The AlexNet [3] is an eight-layer deep network that was the first CNN model used in the deep learning object detection task. It is followed by VGG that uses smaller (3×3) convolution filters instead of 5×5 and 7×7 filters previously used in AlexNet. Then the GoogLeNet [37] increases the width and depth of CNN and, at the same time, introduces the factorizing convolution and batch normalization. In 2015, the ResNet [34] was proposed to simplify the network training phases by reconstructing the network layers and using the reference layer input to learn the residual function. The literature helps the backbone selection in this thesis.

Two-stage

In the literature on object detection methods, the two-stage methods appeared earlier than the one-stage ones. Two-stage detection is also implemented by a region-based convolutional neural network. Considering that the deep learning network could learn more robust, generalised and higher-level feature representations from input volume, Girshick et al. improved the object detection by combining the proposed Regions with CNN features (R-CNN) [18].

R-CNN: The idea behind R-CNN [18] is simple. It starts with creating the region of interest (RoI) by selective search [38]. Because CNN requires all input volumes to have the same size, each image will be cropped by a set of RoI and resized to the same. Then CNN is used to extract the feature representations for classification. R-CNN has made significant progress, but the disadvantages are obvious: feature extraction of over 2000 bounding boxes from one image leads to a prolonged detection in GPU. SPPNet [19] overcomes this problem by generating the same size of pooling result from the various size of inputs.

SPPNet: He et al. proposed Spatial Pyramid Pooling Networks (SPPNet) [19] in 2014. The aim is to address the issue that CNN requires the fixed input (e.g., AlexNet [3] requires a 224×224 image as input). SPPNet allows CNN to generate a fixed-length representation for images of different sizes. Applying it to an object detector (e.g., R-CNN [18]), the backbone can extract features from all proposed bounding boxes at the same time. Compared with the previous R-CNN, it is more than 20 times faster without sacrificing any detection accuracy. Based on SPPNet, Faster R-CNN [11] was proposed in the next year. It takes advantages of SPPNet, proposing Region of Interest Pooling (RoI Pooling) to reduce the training time and achieve state-of-the-art performance.

Fast R-CNN: In 2015, inspired by SPPNet, Girshick proposed Fast R-CNN detector [11]. With RoI Pooling, the detector and bounding box regressor could be trained in the same framework. The mAP increases to 70% from 58.5%, while the detection speed is over 200 times faster than R-CNN. Although the improvement is encouraging, proposal generation is still the bottleneck of detection speed. The reason is that Fast R-CNN relies on the selective search for region proposal while the algorithm is plodding and breaks the entire framework into two parts. Later, Faster R-CNN [11] solves the problem and becomes the first end-to-end detector.

Faster R-CNN: Shortly after the Fast R-CNN, in the same year, Ren et al. proposed Faster R-CNN [11]. It is the first end-to-end deep learning detector and performs near real-time in the detection phase. The main contributions include i) Introducing Region Proposal Network (RPN) to replace the selective search to speed up the proposal generation phase with nearly free cost and ii) bridging the individual region proposal and feature extraction blocks of the Faster R-CNN together to form an end-to-end learning framework. As Faster R-CNN obtains such great success, many algorithms are developed based on or modified from Faster R-CNN, such as Mask R-CNN [39], Libra R-CNN [40], and Cascade R-CNN [41]. Another contribution of Faster R-CNN consists of the translation-invariant anchors and the functions that compute proposals relative to the anchors. Moreover, the proposed method has an order of magnitude fewer parameters to reduce the risk of overfitting.

Feature Pyramid Networks: In 2017, Lin et al. proposed Feature Pyramid Networks (FPN) [21] to address the information loss problem in convolution computation. Before FPN, the detectors take the last feature map of the backbone for bounding boxes regression and classification. Although the last layer benefits category recognition, it is not conducive to localising objects. To fix this problem, the FPN is designed as a top-down architecture to combine high-level features with low-level features to enrich the feature maps. Meanwhile, assigning proposed bounding boxes to corresponding feature maps based on their scale (such as large bounding boxes in high-level feature map while the small ones are in the low feature map) increases the ability to detect small objects. Hence, FPN becomes an essential component of an object detection model.

After a long period of development, Faster R-CNN has demonstrated its powerful ability in object detection tasks. Therefore, Faster R-CNN is to be the baseline in this research.

One-stage

In terms of one-stage detectors, You Only Look Once (YOLO) [22] is the best start because the YOLO series report significant detection speed and comparable accuracy.

You Only Look Once: In 2015, Joseph et al. proposed YOLO. It is the first one-stage detector in deep learning. As indicated by its name, YOLO abandoned the previous ”proposal RoI and detection” paradigm. It applies one network for images. This network uses a grid to split an image, and each of the cells with object centres corresponds to the predicted bounding boxes and probabilities of objects. Meanwhile, YOLO could be considered an anchor-free method because the predicted bounding boxes are not from the predefined anchor. Later, holding the basic idea and inspired by ResNet [34] and SSD [12], Joseph proposed v2 [23] and v3 [24] editions, which further improve the detection accuracy while keeping a very high detection speed. YOLOv4 is proposed by Bochkovskiy et al. in 2020 [25]. The main contributions are verifying the influence of different methods of object detection during the detector training and modifying the cross-iteration batch normalization [42], path aggregation network [43], and spatial attention module [44] for reducing computation cost.

Single Shot MultiBox Detector: In 2015, Liu et al. proposed Single Shot MultiBox Detector (SSD) [12]. The significant difference between SSD and YOLOv1 [22] is that the former one is able to handle the various size of the object because of detecting objects from different feature maps according to the size of the objects, which is the so-called multi-reference and multi-resolution technologies. In contrast, the latter ones only run detection on their top layers.

RetinaNet: In 2017, Lin et al. proposed RetinaNet, which changed the situation that one-stage methods have a significant speed advantage but are far less accurate than the two-stage ones [26]. They claimed that the class imbalance between foreground and background classes is the major cause of accuracy drop. To this end, they proposed "focal loss" to reshaping the standard cross-entropy loss so that the detector could focus more on the challenging, miss-classified examples. Focal loss enables the one-stage detectors to achieve comparable accuracy as two-stage detectors while maintaining a very high detection speed.

Collectively, although the one-stage object detection method has clear advantages in terms of speed and the number of model parameters, the two-stage detection method is better in terms of detection accuracy. The research in this thesis requires higher accuracy than detection speed, so this thesis selects two-stage detectors instead of one-stage detectors.

Anchor-Free

The anchor box concept comes from region proposal network (RPN) in Faster R-CNN, which generates RoIs based on three scales and three aspect ratios, yielding nine anchor boxes. The anchor boxes are considered to violate human intuition to recognise targets, and this is the reason for proposing anchor-free methods.

CornerNet: In 2018, Law and Deng proposed CornerNet [27], which is the first anchor free detector. Instead of using anchor boxes as regression reference, they use two prediction models after the backbone to predict top-left corners and bottom-right corners. They use Hourglass Network [45] as the backbone. For each detector, the outputs are heatmaps for corner detection, embedding for grouping top-left corner and bottom-right corner, and offsets to adjust corresponding corner locations. On the COCO dataset, CornerNet outperforms all one-stage detectors and achieves results competitive to two-stage detectors. In 2019, H. Law et al. updated CornerNet as CornerNet-Lite [46]. CornerNet-Lite is 6x faster than CornerNet with a 1% AP increase, and it is faster and more accurate than YOLOv3.

CenterNet: There are two networks named CenterNet. However, the same name does not mean the same methods. Duan et al. [28] take advantages of cascade corner pooling and centre pooling to output two corner heatmaps and a centre key-point heatmap, respectively. The result is comparable with the state-of-the-art two-stage and one-stage methods on the COCO test-dev dataset. Nevertheless, it is not the only method to detect the object centre. Zhou et al. proposed the other CenterNet [29] in 2019. They treat objects as centre points only. With the modified focal loss, it solves background and foreground imbalance. Meanwhile, it cooperates with offsets computation to reduce discretisation error caused by the output with stride and L1 loss for size regressor. The CenterNet also achieves the state-of-the-art performance.

FCOS: Fully Convolutional One-Stage Object Detection (FCOS) [31] uses another method to avoid using anchor boxes. They directly regress the target bounding box for each location, meaning that the detector directly views locations as training samples. The regression targets become the distances of pixels between the detected result and the ground truth. Meanwhile, they use center-ness, which is the location deviates from the object centre, to down-weight the scores of bounding boxes far from the centre of an object, which improves the detection performance remarkably in the final non-maximum suppression process. FCOS outperforms the anchor-based counterpart RetinaNet by 1.9% and 1.3% in AP with ResNet-101-FPN and ResNeXt-32x8d-101-FPN, respectively. FCOS also outperforms the recent anchor-free one-stage detector CornerNet with much less design complexity.

FoveaBox: In 2019, Kong et al. [30] proposed FoveaBox. It jointly predicts the objects' centre areas and bounding boxes without a predefined anchor. The most interesting point is that they generated a positive area (named object fovea) from the corresponding ground truth bounding box. The object fovea is a small area in the object centre instead of an object centre pixel. Outside the fovea, the model takes the area as negative area. This model is a state-of-the-art model on the COCO test dataset.

Transformers: In 2020, Carion et al. proposed End-to-End Object Detection with Transformers [32]. It is the first to use transformer [47] technology in the object detection area. Moreover, its architecture is much simpler compared with previous detection architecture (e.g. Faster R-CNN). The idea behind it is simple: it aims to train N object queries to find where the object is, and what they are, where the N is predefined as the number of objects in one image. Technically, similar to other architectures, they use CNN as a backbone to extract features from images. Cooperating with Position Embedding, the feature map is then reshaped to have a size of $WH \times C$ (representing image width, height and the number of channels, respectively) in order to be fed into the transformer encoder sequentially. The encoder features are sent to the transformer decoder, who computes N object queries as a set of predictions. The most instructive part is that they use an attention matrix to locate objects. Transformer [47] enables communication between different parts of the image. Hence, in the attention matrix, which has the size of $WH \times WH$, each point corresponds to two pixels in the image to describe their relationship. A simple architecture like this outperforms Faster R-CNN. However, it requires 8 GPUs to train for six days at the COCO dataset and the performance drops when detecting small objects.

Overall, the detection methods without the anchor setting show how to eliminate the dependence on anchor in the regression process and provide an example of transforming the regression target.

Deep Learning Applied on Remote Sensor Applications

In this thesis, all the collected images are taken by a drone, so it is necessary to step into the remote sensor detection area.

As introduced before, the detection models are categorised into one- and two-stage methodologies. Similar to the common object detection tasks, those methods also achieve considerable success in the task of detecting objects on drone images, such as vehicle detection and human detection. According to Mittal et. al [48], Faster R-CNN and YOLO are the most popular methodologies as they give the best detection accuracy and have the highest detection speed, respectively.

Additionally, deep learning applications using remote sensor (e.g. satellite and drone) data have a different nature from those in common object detection datasets. That is, the objects in remote sensor images are usually much smaller in size than those in common object detection datasets. In 2018, Kellenberger et al. presented work in detecting deer in UAV images [49]. They take 512×512 images as input and output a 32×32 probability score to indicate if such an area contains animals or not. In the following year, Pang et al. [50] proposed R^2-CNN to detect objects in large-scale remote sensing images. The R^2 means remote sensing and region-based. This work can be considered as an extension of Faster R-CNN. The extension is inserting global attention after feature extraction.

The significance of these works are i) they give a clue to deal with high-resolution images (e.g. crop them into small patches), ii) the models are robust to false positives and strong to detect tiny objects, and iii) their results show that a shallow network may be sufficient in the tasks of detecting objects from drone or satellite images.

2.3 Chapter Summary

In this chapter, the review emphasises the importance of constructing a microatoll dataset for microatoll detection from drone images. The literature of the existing datasets shows that microatolls are not considered in the common datasets used in object detection and drone image based detection challenges. Therefore, constructing a reliable dataset is the critical step for the research work in this thesis. In order to enhance the reliability of annotation in our microatoll dataset, the literature is reviewed to establish a labelling rule for microatolls in drone images.

After that, through LeNet, the pioneer of convolutional neural networks, this chapter gives an overview of critical layers used CNN. Those essential layers lay the foundation of the feature extraction backbones. Those backbones help in building robust deep learning based object detectors. This thesis selects Faster R-CNN (a two-stage method) as the baseline model because the accuracy of detection microatolls is more important than the detection speed. The literature in modern deep learning based object detection shows the two-stage methods have advantages in object detection accuracy compared with the one-stage ones. Besides, the real-world applications of detecting objects (i.e. animals) are reviewed.

Chapter 3

Microatoll-Net For Microatoll Detection

Contents

3.1 Introduction

The capacity to detect individual corals from airborne remote sensing imagery with object detection is an exciting technological frontier, with the potential to deliver a wealth of important information for reef management [51]. In practice, it is complicated and inefficient for researchers to survey an island to locate microatolls physically. The use of drones has recently become a more efficient solution. Particularly, as drone platforms coupled with GPS technology enable images to be georecitified, such that features identified in images can be located in a geographical space, i.e. across different sub-zones of reef platforms. Accordingly, automatically detecting microatolls from the captured high-resolution images becomes a key step of this solution.

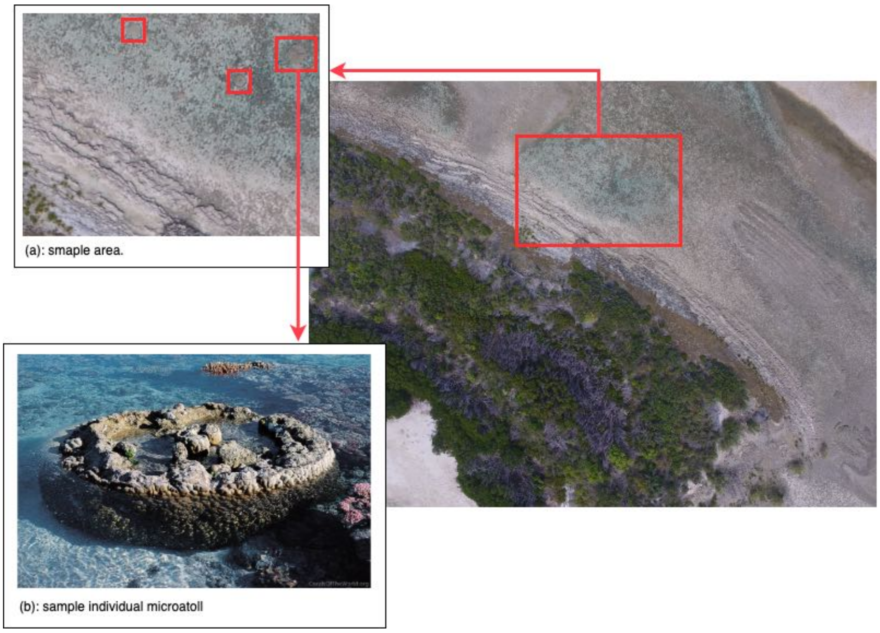


Figure 3.1: This figure shows the situation of microatoll detection. Subfigure (a) is zoomed into an image of an island captured by a drone. All targets in (a) are indicated by red boxes. Subfigure (b) gives an example of an individual microatoll.

The challenge of detecting objects from drone images has recently attracted research attention [52]. Although popular object detectors have achieved great success in common objects detection settings (i.e. VOC [8] and COCO [9] challenges in literature), they cannot be simply transplanted to detect objects viewed from drones. The situation is particularly challenging in microatoll detection presented in this chapter. As seen in Fig. 3.1, a microatoll is a tiny object in drone images. Moreover, it could be fully covered by water, and its appearance is affected by the motion of water. Microatolls could also be confused with other corals nearby, and dead and living microatolls could connect with each other, forming a flat, coalesced pavement that often misleads object detectors.

In this chapter, by accessing a set of real images of an island collected by a drone, the nature of this microatoll detection task is examined. Through the joint work from the fields of computer science and coastal geography, we manually label the ground truth of individual microatolls and iteratively refine the labels by leveraging the detection result of a trained detector. This yields a quality benchmark dataset that can reliably be used to train a deep learning based microatoll detection model and evaluate its performance. Focusing on the commonly used Faster R-CNN [11] detector, the efficacy of visual object detection for this microatoll detection task is investigated. Firstly, an investigation is conducted on the fine-tuned Faster R-CNN with online hard example mining (OHEM) [53] and feature pyramid networks (FPN) [21]. Then, the non-maximum suppression (NMS) threshold is adjustified to improve detection precision, and finally, promising detection results are achieved. Meanwhile, OHEM and the adjusted NMS threshold reduce the recall and average precision, and there are still considerable false positives appearing around microatoll pavements, which affects the overall detection performance.

To address this situation, the characteristics of individual microatolls shall be better reflected in the detection model. The comparison between individual microatolls with distractors shows that the individual microatolls are usually surrounded by water and that this does not apply to the latter. This characteristic motivates us to develop the Microatoll-Net explicitly for the microatoll detection task. It has a novel module called context region of interest to actively extract the information on the surrounding area of each bounding box proposal to improve discrimination. Several ways are designed to utilise this information, either directly combining with original features or working as a signal via an attention mechanism to weight original features.

Experimental study is conducted on the created microatoll dataset. The proposed Microatoll-Net demonstrates the overall best detection performance. The most significant improvement is obtained on the most challenging region for detection, indicating the efficacy of the proposed context region of the interest module.

3.2 Related Work

This chapter aims to extract contextual features to help microatoll detection from drone images. In the literature, some work has used context to boost detection. For example, Bell et al. [54] use IRNN [55] to extract contextual information while the IRNN is a set of recurrent neural networks, introducing more computational cost. PyramidBox [56] utilises the head and body information to facilitate face detection. Besides, Kuan et al. [57] propose region average pooling (RAP) to extract the contextual information on object co-occurrence in an image. Specifically, RAP utilises all regions of interest (RoI) to compute an average representative and concatenate it with each RoI, respectively, for object detection. Note that IRNN and PyramidBox are applied to a one-stage detector while the RAP is designed for Faster R-CNN. There will be an experimental comparison of our context-RoI module with RAP for microatoll detection.

Meanwhile, second-order pooling (SOP) technical attracts us in modelling feature relationships. Wang et al. [58] highlight that the feature representation from kernel matrix is a more sophisticated feature relationship in object classification, and Gao et al. [59] utilise SOP to generate an attention signal from the feature map as a weighting vector, and it achieves significant performance. Accordingly, it is presumed that modelling the relationship between context and object feature would improve the performance. By contextual features, this thesis means the features presented in the surrounding area of an object to detect. Hence, an experimental study is conducted to check whether the SOP context feature or ordinary CNN context feature can improve detection performance.

3.3 Microatoll Dataset Construction

Deep learning is a data-driven technique. The quality of data significantly affects the design, training and final performance of a deep learning model. Without being properly processed, data could mislead the model, and the model will not be able to learn the underlying patterns from it. This section presents how the microatoll dataset is collected, organised and processed to meet the requirements of being reliably used in this microatoll detection task.

3.3.1 Microatoll Data Collection

The microatoll photographs are collected by a drone from Nymph island, a national park in Far North Queensland, Australia. The Nymph island is in 1,636 km northwest of Brisbane, and located at $14°39'20''S, 145°15'03''E$. The island is circular, and there is a lagoon in the centre.

The fieldwork was conducted to scan the whole Nymph island to locate the individual microatolls for further research. Because the island area is around 0.65 km^2, using a drone could efficiently reduce working hours and safely store high-quality images. To prepare for the drone surveys, a DJI Phantom drone was adopted, and the flight route was predesigned to scan the whole island at an altitude around 60 to 70 meters. Finally, 1,400 drone images were acquired in total.

3.3.2 Microatoll Data Processing

Initially, a labelling software [60] is used to generate the ground truth bounding boxes for each image manually. The annotations format follows the VOC [8] to use an XML file to store the top-left and button-right corners of the ground truth bounding boxes. Before labelling microatolls, we need to locate the areas that may contain microatolls roughly. It is notable that, first, all images have the same resolution because they are collected by one drone at the same height. Second, an image may overlap with another image by about 60% of its size because of the need for geological picture collection, which means an area of the island presented in an image could appear multiple times in other images. Meanwhile, the overlapping reminds us of the potential data leakage issue.

According to Scoffin and Stoddart [2], all living microatolls grow in the open sea, and their tops are higher than the lowest water level of spring tides. Furthermore, they indicate that the positions containing microatolls are around $14°39' lat.S, 145°15' long.E$. Benefited from the literature, we identify 195 images that contain microatolls to construct the dataset for this microatoll detection task.

This dataset should be able to be split into a training and a testing subset. The former is used to train an object detection model, while the latter is used to evaluate the detection performance of this model. Note that there shall be no overlapping between the training and test subsets, and the evaluation result will be over-optimistic otherwise (i.e. the over-fitting issue). Nevertheless, the overlapping between images mentioned earlier makes the splitting not straightforward. Randomly splitting all images into a pair of training and test subsets will not work.

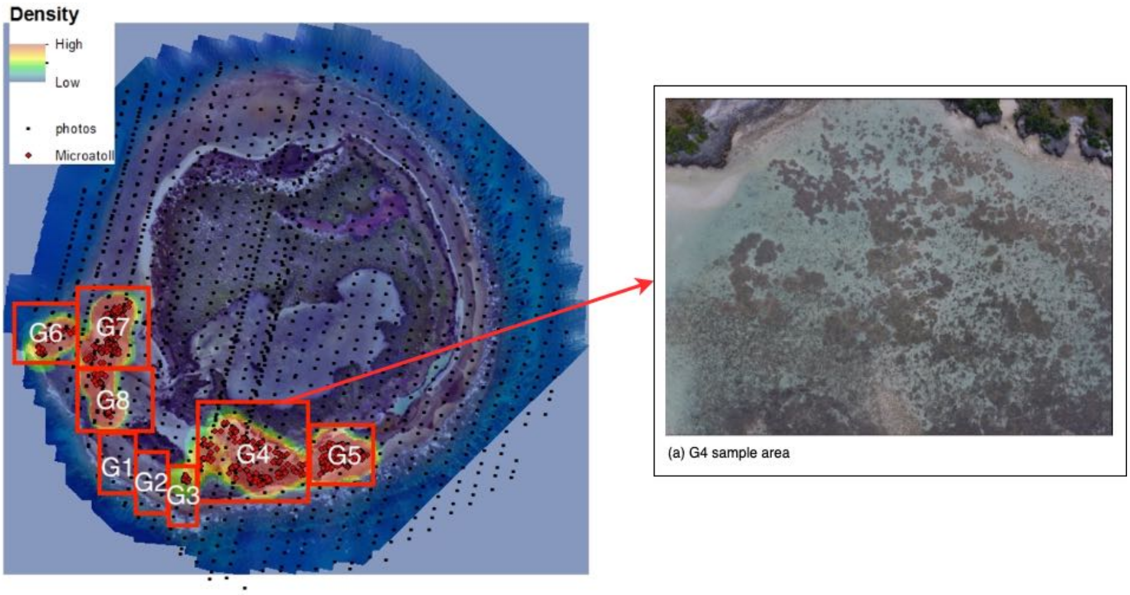


Figure 3.2: The eight groups (from G1 to G8) are formed by the images based on their coverage areas. The colours from blue to red indicate the density of microatolls in an area, and the black and red dots indicate the positions where images are captured, and microatolls are found. Group G4 is the most challenging area because it contains a lot of microatoll pavements shown in the zoomed-in image.

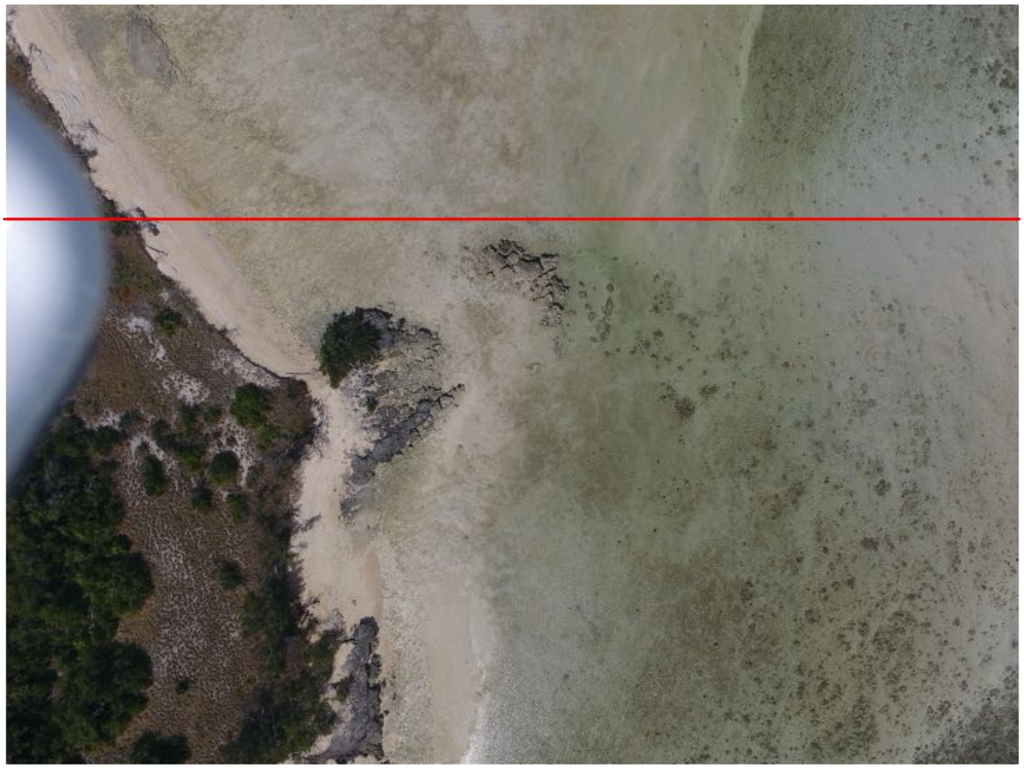


Figure 3.3: An example of separating an image based on an boundary.

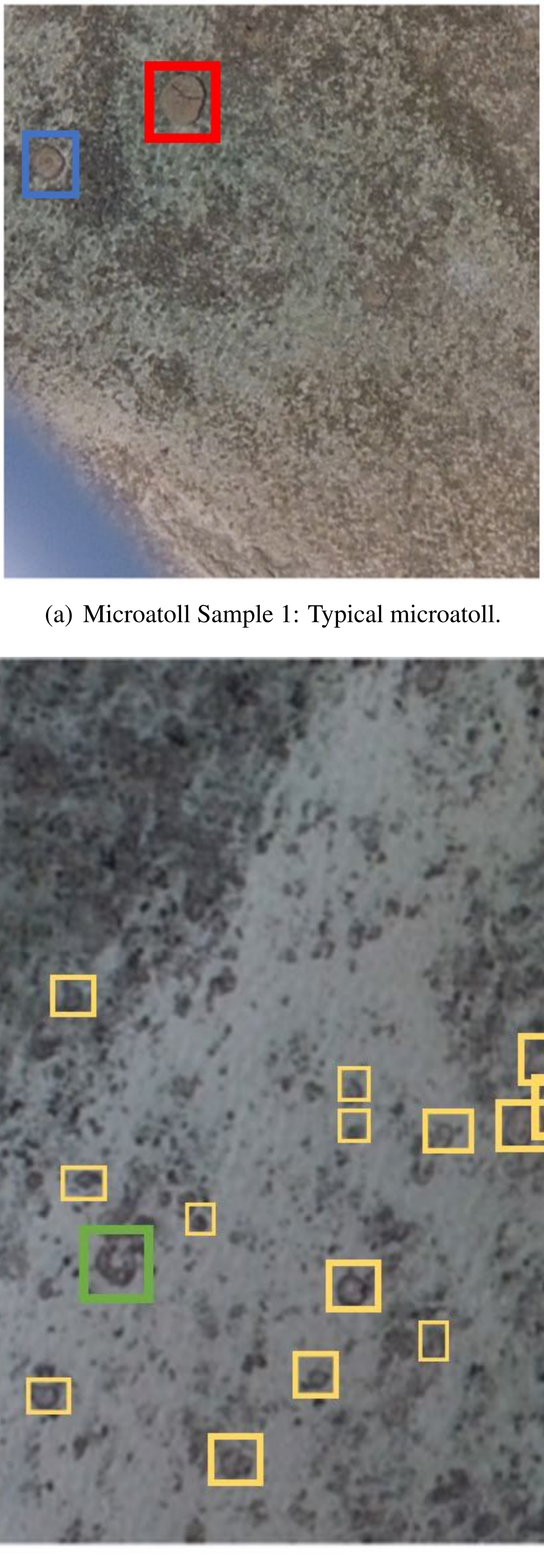
As aforementioned, one individual image could be overlapped with others, and images in this situation should be partitioned. Fig. 3.3 shows an example of how the partitioning is done. The red line is selected according to the gap between areas, and it indicates the separation of the image into two parts. The part above the red line belongs to one area, while the part below the red line belongs to another area. In this way, all the 195 images in the dataset are examined and processed. After this procedure, images corresponding to eight groups of the areas (i.e. G1, . . . , G8) that contain microatolls are obtained.

3.3.3 Microatoll Data Labelling

Manually labelling data is to create the ground truth for a machine learning model to learn. In this chapter, the target is a bounding box that delineates the location and size of a microatoll to detect. The labelling needs to deal with some difficulties. First, there are no specific rules to indicate what a microatoll shall look like in the images captured by a drone. Second, microatolls are generally very small in size when compared with the size of a whole image. Hence, it is hard for human observers to locate the microatolls even from an area that may contain microatolls.

The research of Scoffin and Stoddart [2] and that of Smithers and Woodroffe [1] indicate that microatolls are commonly in shapes of circles or hose-shoes. Meanwhile, their edges will be significant when compared with the dead top. Based on these two observations, the individual microatolls in drone images are described as follows: in an area of open water and water in moats, a rough circle with a margin darker than its centre could be a microatoll. Also, a microatoll could be with damage or of being a specific horse-shoes shape. After manually labelling, an object detector is trained to work on all images. The top-ranked detection results are included in the ground truth once confirmed to be microatolls. This iterative approach effectively improves labelling efficiency and accuracy.

With the above rules, manual labelling is still time-consuming. Each image in the dataset has around 4000×3000 pixels ($160m \times 160m$), but the size of a microatoll bounding box is in the average of 50×50 pixels ($2m \times 2m$). This size gap could easily lead to mislabelling. The mislabelling means that some microatolls could be missed because of



(a) Microatoll Sample 1: Typical microatoll.

(b) Microatoll Sample 2: Different in size and shape.

Figure 3.4: Microatoll Samples. Those images give the samples in the dataset. The blue rectangle indicates a typical microatoll with a complete shape. The red one shows a microatoll with breakage. The green one is a microatoll in the horse-shoes shape. The yellow boxes indicate the microatolls of other types and some of them are found via the iterative labelling process.

their small sizes and that the microatolls with low confidence are not labelled. Minimising the odds of mislabelling is crucial to ensure the high quality of ground truth. The iterative labelling process also addresses this issue to enhance reliability. In addition, the dataset is confirmed by human experts, and all annotations are adjusted accordingly.

The setting for training and testing an object detection algorithm is described as follows. The size of an input image is set as 800 pixels in width and 600 pixels in height. This setting is based on the following considerations. On one hand, the size of microatolls is in the range 25×25 pixels ($1m \times 1m$) to 125×125 pixels ($5m \times 5m$). Using 800×600 pixels ($32m \times 24m$) input image will contain enough background information for the object detector to use. On the other hand, further increasing the input image size will incur more computational cost for object detection algorithms.

To facilitate training, there are 1888 training patches in the size of 800×600 pixels by sampling the original large images (in the size of 4000×3000 pixels as mentioned before) with the stride of 400 and 300 pixels. Choosing this stride setting is to balance the number of generated training samples and the area of overlapping regions between two adjacent training samples.

It is worth mentioning that the labelling results will continue to be adjusted during developing and applying the object detection algorithm. This process takes advantage of the detection algorithm (which could find some microatolls missed in the human labelling process) to enhance the labelling quality. Fig. 3.4 gives an example image in the dataset, in which the coloured bounding boxes highlight the labelling result. Finally, it takes about 150 hours to label 7,414 annotations for all the 195 big images (4000×3000 pixels).

3.4 Model Investigation And Analysis

This chapter is to detect microatolls from given images collected by a drone automatically. Considering that i) the one-stage detectors could generate more false-positive samples than the two-stage ones, and ii) there is currently no need for high-speed detection, this investigation adopts the two-stage (Faster R-CNN) object detection algorithm.

3.4.1 Investigation Setting

All experiments are conducted in the same server with 12 Intel(R) Xeon(R) CPU E5-2603 v4 @ 1.70GHz, 31888 MB memory and 4 GeForce GTX1080 GPU cards. This experimental study investigates the performance gap between SSD (i.e. One-stage detector) and Faster R-CNN (i.e. Two-stage detector) on the constructed microatoll dataset. We compare them because they are the typical representatives of their respective approaches. The comparison is conducted based on their default structures and settings.

3.4.2 Evaluation Method

The following criteria are used to evaluate the detection performance. The ground truth means the labels (i.e. the bounding boxes manually labelled for each microatoll) in the microatoll dataset. The detected number means the number of microatolls successfully detected by a model. A successful detection means that the intersection over union (IOU) of the ground truth bounding box and a detected bounding box is greater than a predefined threshold (usually 0.5). The IOU between the ground truth and detected bounding boxes is calculated as

$$IOU = \frac{\text{area of overlap}}{\text{area of union}} \tag{3.1}$$

The true positive means that a microatoll is correctly detected, and the false positive means a non-microatoll area is wrongly detected as a microatoll. The recall indicates the percentage of successfully detected microatolls in all the microatolls presented in an image, and the precision indicates the percentage of the true microatolls in all the detected ones. Average precision (AP) is calculated from recall and precision to give an overall evaluation of the detection performance of the model. Generally, AP equals the area under the precision and recall curve. The AP is calculated as follows, and it finally presents the detection score (higher is better).

$$AP = \int_0^1 p(r)dr, \tag{3.2}$$

where $p(r)$ is the precision that takes recall as a parameter.

3.4.3 Investigation Result

Table 3.1 and Table 3.2 give the evaluation results on the detection methods of SSD and Faster R-CNN, respectively. Considering that the training and detection time cost is different between one-stage and two-stage detectors, the comparison only focuses on their performances on the testing dataset. Recall that the microatoll dataset images correspond to eight areas (G1 to G8). Groups 1, 2 and 3 are merged into one due to their smaller sizes and their closeness. Consequently, there are six area groups G123, G4, G5, G6, G7 and G8. The leave-one-out cross-validation protocol [61] is used in this experimental study. It reserves one group as the test set in turn and uses all the remaining groups for training.

Comparing the results in Table 3.1 and Table 3.2 indicates that the two-stage detector Faster R-CNN is more promising because it achieves higher average precision than the SSD method. With the leave-one-out cross-validation protocol [61], this evaluation result should be convincing. As aforementioned, the dataset consists of six groups (i.e. G123, G4, G5, G6, G7 and G8) without any image overlapping among these groups. The results obtained by testing the trained detection model on each unseen group indicate the generalisation capability of the model. The significant gap between SSD and Faster R-CNN in terms of AP is because SSD is trying to use more bounding boxes to cover the microatolls (therefore, precision is lower). The low precision means that SSD is weaker in distinguishing similar background regions from microatolls. Based on this observation, Faster R-CNN is employed as the baseline method to detect microatolls to produce better detection performance. (ResNet50 [34] as default backbone if not mentioned.)

3.4.4 Performance and Analysis

Fig. 3.5 illustrates the path to obtain the final result. It gives an overview of this section as well. Following the path, all the details will be given. Faster R-CNN gives a better

Model	Tested Group	Ground Truth	Detected	True Positive	False Positive	Recall	Precision	Average Precision
Basic_SSD	G123	655	23600	618	22982	0.94	0.03	0.72
	G4	7853	110400	6964	103436	0.89	0.06	0.44
	G5	502	14200	490	13710	0.98	0.04	0.67
	G6	3297	82800	3109	79691	0.94	0.04	0.69
	G7	4553	82200	4458	77742	0.98	0.05	0.78
	G8	3387	63800	3185	60615	0.94	0.05	0.63
Average						0.945	0.045	0.66

Table 3.1: Basic SSD model performances in the microatoll detection. “Tested Group” indicates the group used as the test set, while the remainder is used together as the training set.

Model	Tested Group	Ground Truth	Detected	True Positive	False Positive	Recall	Precision	Average Precision
Basic_Faster_R-CNN	G123	655	1872	628	1244	0.96	0.34	0.87
	G4	7853	24406	7104	17302	0.91	0.29	0.62
	G5	502	2144	476	1668	0.95	0.22	0.76
	G6	3297	10328	3163	7165	0.96	0.31	0.84
	G7	4553	13950	4440	9510	0.98	0.32	0.87
	G8	3387	9366	3256	6110	0.96	0.35	0.83
Average						0.953	0.305	0.79

Table 3.2: Basic Faster R-CNN model performances in the microatoll detection. “Tested Group” indicates the group used as the test set, while the remainder is used together as the training set.

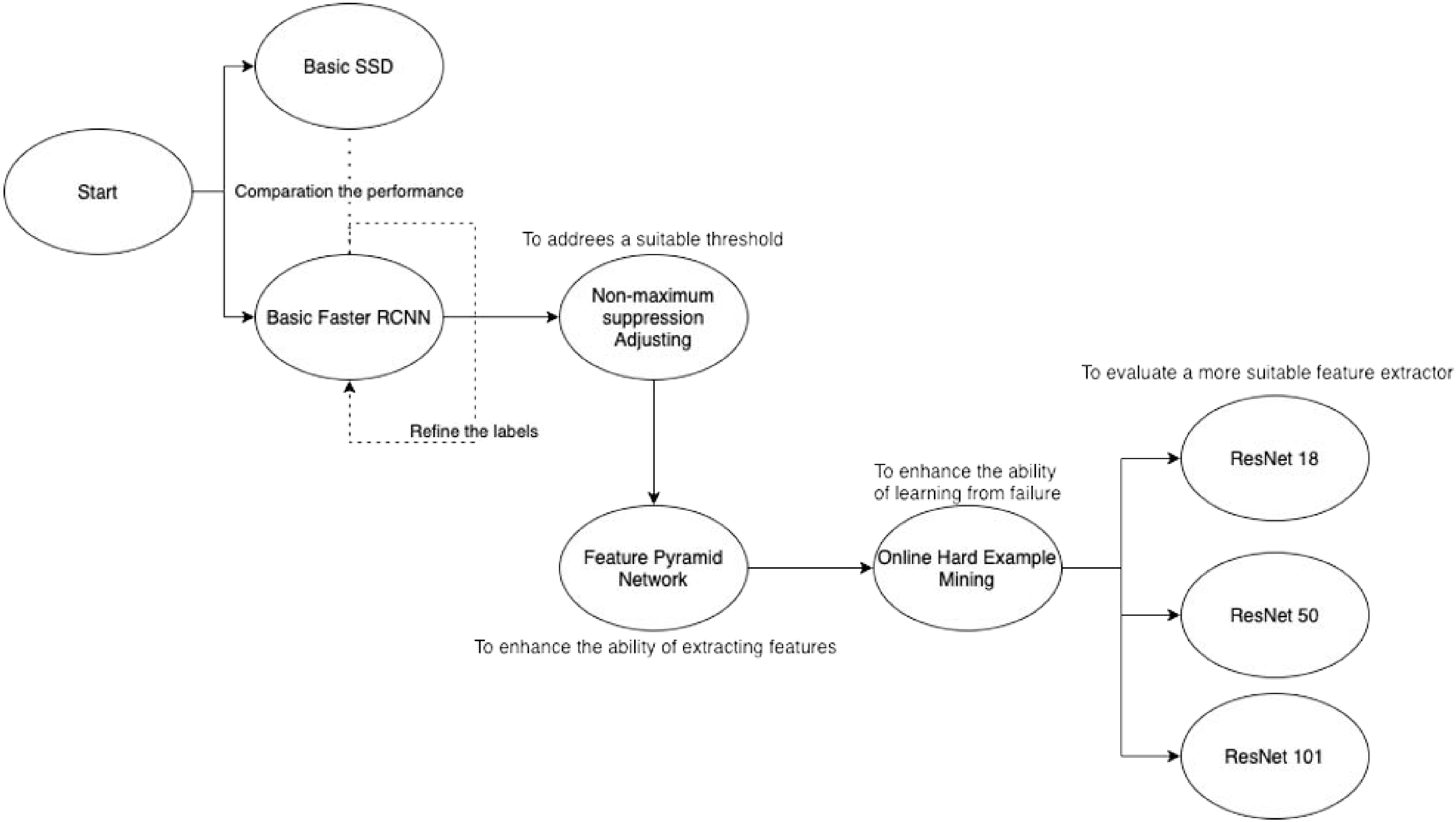

Basic SSD
Start
Comparation the performance
To addrees a suitable threshold
Basic Faster RCNN
Non-maximum suppression Adjusting
Refine the labels
To evaluate a more suitable feature extractor
ResNet 18
To enhance the ability of learning from failure
Feature Pyramid Network
Online Hard Example Mining
ResNet 50
To enhance the ability of extracting features
ResNet 101

Figure 3.5: The path to obtain the final result. This flowchart indicates the path to obtain the final result. It starts from the comparison of SSD and Faster R-CNN, then obtains the improvements step by step. And finally, it is determined that a deeper backbone may not be necessary to boost the performance.

result than SSD, but its performance is still far from being satisfactory. The low precision means many detected microatolls are actually from the background area of an image. At the same time, reef flats (the environment where the microatolls are located) are very dense and heterogenous parts of the image/ reef environment. To overcome this problem, adjusting the NMS parameter becomes the first and most straightforward option, as explained below.

As observed from the experimental study, a lower IOU threshold at the testing phase will keep more detection results (even if they are false positive). Based on the AP formula, the area under the precision and recall curve is more significant. However, more false positive detection results may require more human resources to distinguish and remove. The NMS algorithm developed in the literature is designed to only detect an object once. Hence, those bounding boxes overlapped with one target will be removed while the highest confidence score will remain.

The confidence score for each detection box in the detection phase indicates how probable it is indeed the target object. To obtain higher confidence instead of all of the bounding boxes, increasing the confidence threshold at the test phase is essential. With IOU threshold and confidence thresholds set as 0.3 and 0.5 at the testing phase, respectively, the improved results in all groups are obtained as follows.

The comparison between Table 3.2 and Table 3.3 shows that the precision values are well improved with the mild decrease of recall value in all regions, except G4. The experiment result demonstrates the effectiveness of the above strategy of NMS adjustment. The following issues remain:

1. What leads to the inferior detection performance on the group G4?

2. How to further boost the detection precision?

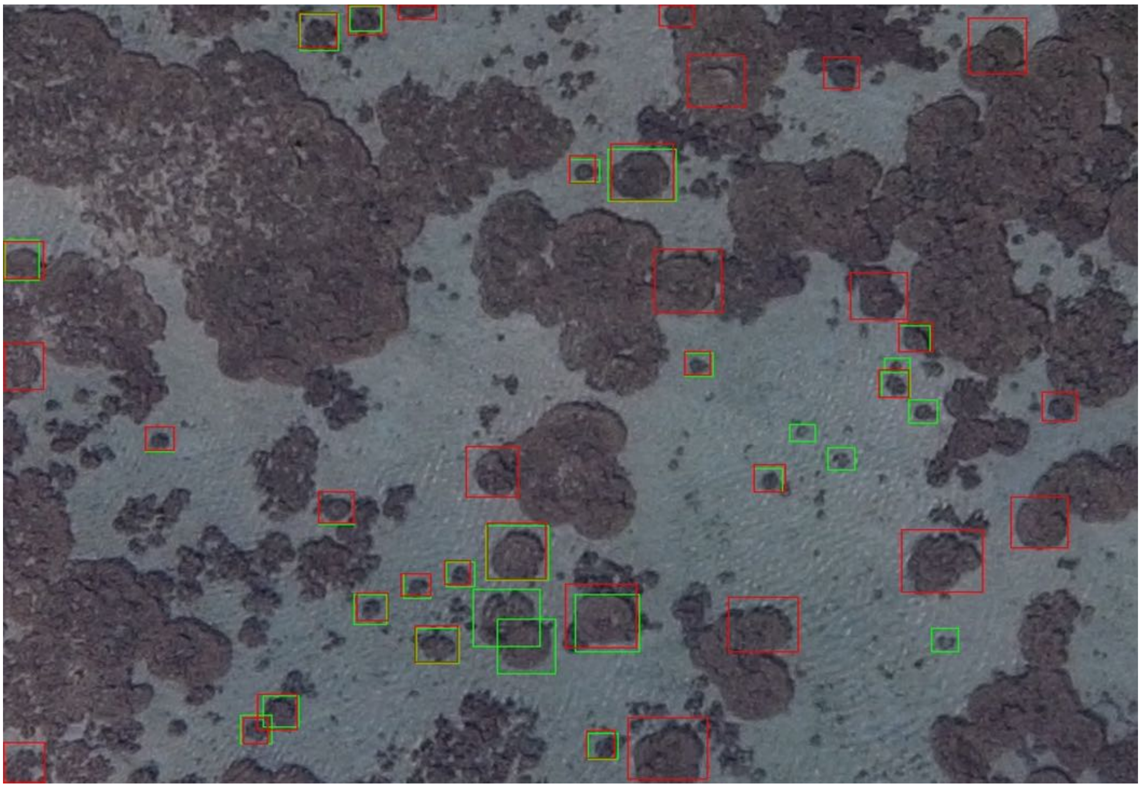


Figure 3.6: An example detection result from the group G4. The green bounding boxes indicate the ground truths while the red ones are the detection results. Meanwhile, it shows that the not only the microatolls vary in size, and also the microatoll pavements are in different size and shape.

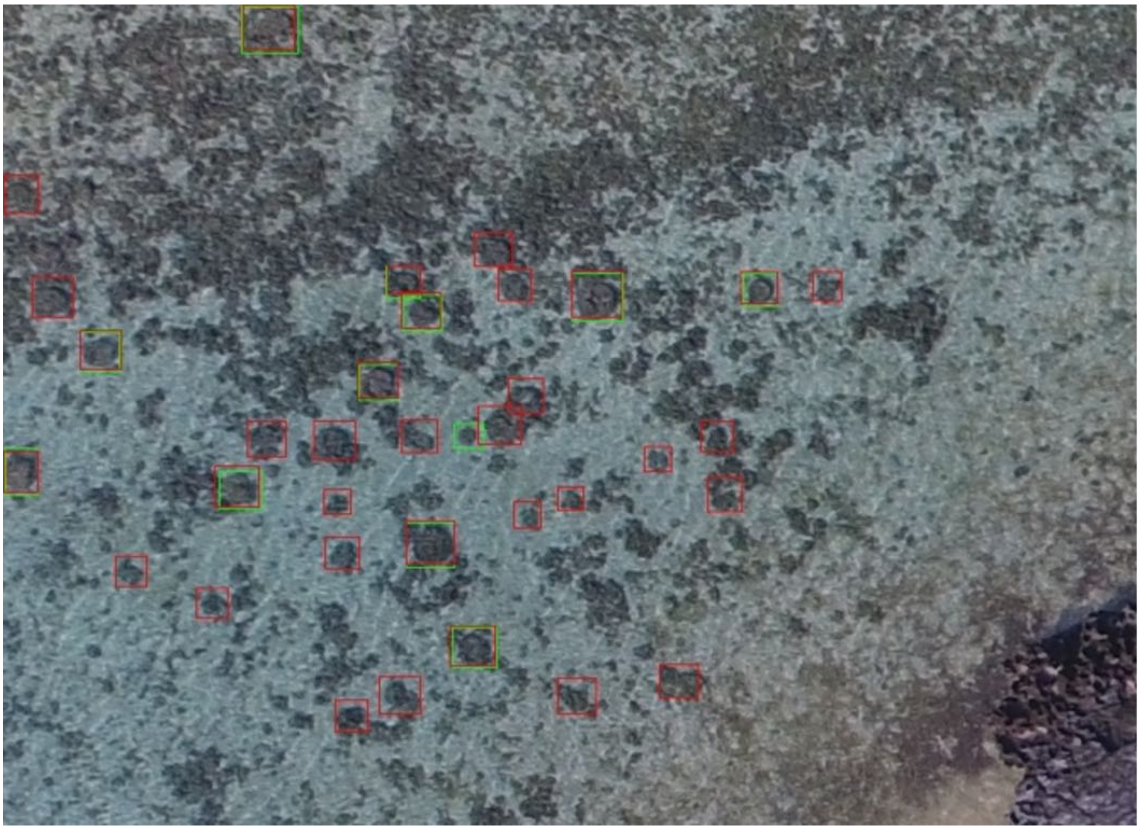


Figure 3.7: One detection example of group G5. The green bounding boxes indicate the ground truths while the red ones are the detection results.

Model	Tested Group	Ground Truth	Detected	True Positive	False Positive	Recall	Precision	Average Precision
Basic_Faster_R-CNN With NMS adjustment	G123	655	943	584	359	0.89	0.62	0.84
	G4	7853	10502	5542	4960	0.71	0.53	0.54
	G5	502	1085	454	631	0.90	0.42	0.75
	G6	3297	5030	2911	2119	0.88	0.58	0.81
	G7	4553	7353	4208	3145	0.92	0.57	0.85
	G8	3387	4676	2927	1749	0.86	0.63	0.78
Average						0.86	0.56	0.76

Table 3.3: After changing IOU and confidence thresholds to 0.3 and 0.5, respectively, the increase of Precision is obvious.

Model	Tested Group	Ground Truth	Detected	True Positive	False Positive	Recall	Precision	Average Precision
Basic_Faster_R-CNN With FPN	G123	655	860	580	280	0.89	0.67	0.84
	G4	7853	8181	5167	3014	0.66	0.63	0.55
	G5	502	921	460	461	0.92	0.50	0.79
	G6	3297	4799	3000	1799	0.91	0.63	0.85
	G7	4553	7205	4301	2904	0.95	0.60	0.88
	G8	3387	4218	2964	1254	0.88	0.70	0.81
Average						0.87	0.62	0.79

Table 3.4: Basic Faster R-CNN obtains improvements by combining the sub-neural network FPN. The Recall becomes slightly higher while the Precision achieves clear improvement over the results in Table 3.3.

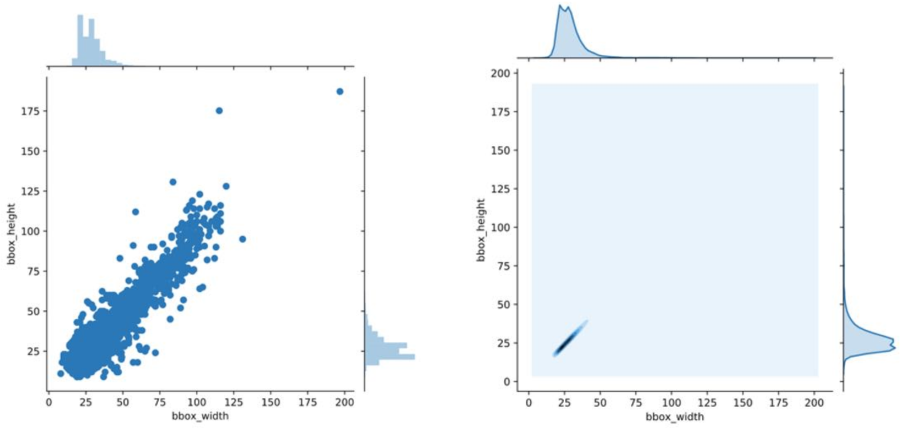
Fig. 3.6 illustrates the issues in detecting microatolls from the group G4 by showing a typical situation. Firstly, the microatolls presented in the images in this group have various sizes. Secondly, the microatolls here grew and connected, forming microatoll pavements. Those two factors lead to the missing detection of microatolls and a large number of wrong detections.

Additionally, Fig. 3.7 shows another reason in the area G5 that causes low precision. As seen, many small regions in this image look similar to microatolls. However, they are actually not—those results are false-positives.

To sum up, the main factor causing low recall is that some microatolls are too small to be detected, and the main factors leading to low precision are the pavements formed by connected microatolls and the backgrounds that are similar to microatolls.

According to the above result, it can be seen that the detection model shall learn more discriminative features to distinguish similar background regions from microatolls. Also, the detection model could focus more on the bounding boxes with the expected size of microatolls. This can be achieved by using the feature pyramid network (FPN) as a sub-neural network. To estimate the expected size of microatolls, a statistical analysis of the ground truth bounding boxes is conducted as follows. In total, there are 20,242 ground truth bounding boxes labelled in the microatoll dataset. As shown in Fig. 3.8, the height and width of most bounding boxes, which are taken at the same altitude around 70m, distribute in the range from 25 pixels to 75 pixels (as indicated by the two histograms in Fig. 3.8(a) and the Fig. 3.8(b)), and the ratio of height to width is roughly 1.0.

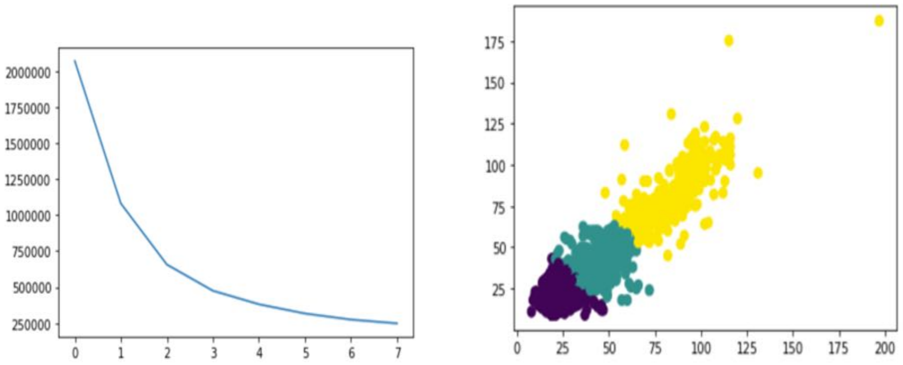
We use clustering method to find the suitable anchor setting for the model. K-means [62] clustering algorithm is used in this experiment to discover the underlying clusters of the bounding boxes. In other words, the K-means is to cluster all the bounding boxes based on their height and width. As shown in the subfigure 3.9(a), the elbow-point for clustering is 3, which means the bounding boxes can be reasonably clustered into three clusters. The subfigure 3.9(b) colours obtained the three clusters in different colours. We adjust the scale and aspect ratio of the anchor and combine the FPN with Faster R-CNN based



(a) Scatter plot for microatoll bounding box width and height.

(b) Density map for microatoll bounding box width and height.

Figure 3.8: Bounding boxes height and width relationship. The x-axis and y-axis represent the width and height of a bounding box.



(a) Elbow point of selecting K for k-mean algorithm.

(b) Clustering results (3 groups)

Figure 3.9: By using K-means clustering algorithm, all the bounding boxes are clustered into three clusters according to their height and width. The different colour indicates the bounding boxes are clustered into 3 different groups.

Model	Tested Group	Ground Truth	Detected	True Positive	False Positive	Recall	Precision	Average Precision
Basic_Faster_R-CNN With FPN and OHEM	G123	655	665	537	128	0.82	0.81	0.79
	G4	7853	5075	3714	1361	0.47	0.73	0.40
	G5	502	662	418	244	0.83	0.63	0.73
	G6	3297	3511	2695	816	0.82	0.79	0.78
	G7	4553	5511	4015	1496	0.88	0.73	0.83
	G8	3387	3020	2522	498	0.75	0.84	0.71
Average						0.76	0.76	0.71

Table 3.5: This table shows the detection performance after incorporating the OHEM strategy. The increase of Precision is significant when compared with Table 3.4.

Model	Tested Group	Ground Truth	Detected	True Positive	False Positive	Recall	Precision	Average Precision
R101_Faster_R-CNN With FPN and OHEM	G123	655	657	522	135	0.80	0.80	0.77
	G4	7853	5224	3816	1408	0.49	0.73	0.42
	G5	502	661	396	265	0.79	0.60	0.68
	G6	3297	3429	2565	864	0.78	0.75	0.73
	G7	4553	5452	3917	1535	0.86	0.72	0.81
	G8	3387	3053	2399	654	0.71	0.79	0.67
Average						0.74	0.73	0.68

Table 3.6: A deeper backbone model is used for the object detection model. The performance degrades.

on this clustering result. The obtained result is shown in Table 3.4.

Table 3.5 further presents the results of combining train another strategy called online hard example mining (OHEM) [53] into the object detection model. The OHEM can better train the detection model by helping the model focus on the bounding boxes for which it does not function well. Technically, in the forward propagation step of training, the OHEM will sort the region of interest (RoI) by their loss values. After that, in the backward propagation step of training, the OHEM will select the most *K* difficult RoIs to train the model. By taking this strategy, it can be found in Table 3.5 that the precision values are well improved when compared with Table 3.4.

The following experimental study investigates for this microatoll detection task whether the object detection model can achieve better performance by using a deeper CNN neural network as its backbone network. As seen in Table 3.6, the result indicates that using a deeper backbone network (i.e. ResNet101) is not necessarily helpful in increasing the detection performance. The deeper model may cause the performance decrease because it contains more parameters to tune and it is not well trained by the limited number of samples in our microatoll dataset.

To take an overview of these improvements, two diagrams are plotted to show the increase achieved step by step. As Fig. 3.10 illustrates, with a slight loss in recall, the precision increased from around 0.3 to near 0.8. Considering the G4 is more challenging than the others (as the G4 mainly contains the microatoll pavements), it is treated as an outliner. Hence, removing the G4 and obtained Fig. 3.11. The promotion is significant. The recall remains above 0.8, and the precision goes up to around 0.8 from nearly 0.3. More details are in Fig. 3.12 and Fig. 3.13.

By comparing with the result in Table 3.4 and Table 3.5, the improvement in the same area is significant. Specifically, Fig. 3.14 shows the differences between the original setting (corresponding to Fig. 3.6) and the adjusted setting reported in Table 3.5. As seen from the sub-figure (a) of Fig. 3.14, while microatolls are correctly detected, many similar background regions are wrongly detected as microatolls, resulting in a lower precision

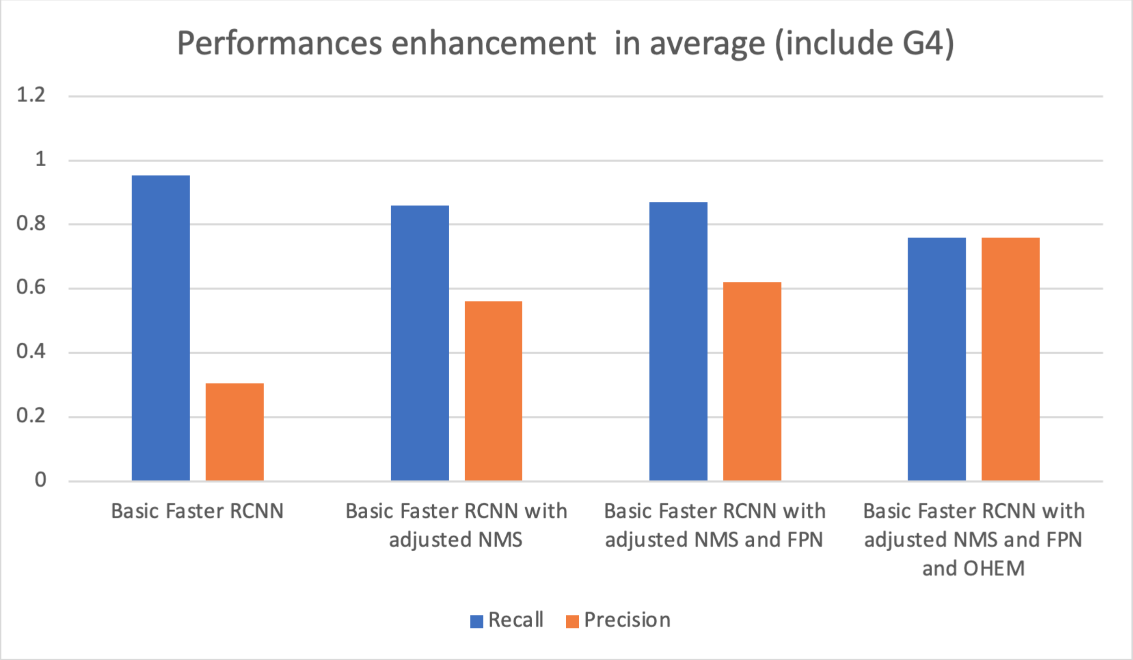


Figure 3.10: Performances enhancement in average (include G4)

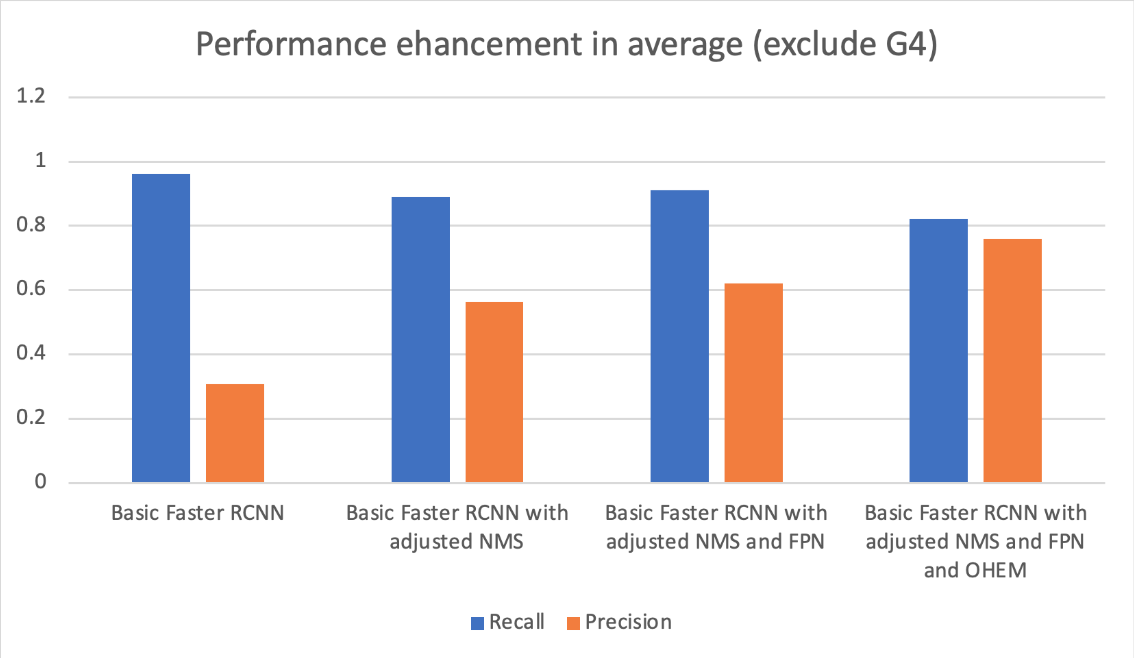


Figure 3.11: Performances enhancement in average (exclude G4)

(this is the setting in Table 3.4). In contrast, the sub-figure (b) of Fig. 3.14 shows that with the same number of microatolls detected, fewer wrong detection results occur, leading to a higher precision. Visually, fewer unpaired red bounding boxes are presented in the right sub-figure (b). The red bounding boxes paired well with the green boxes indicate that the ground truths are correctly detected.

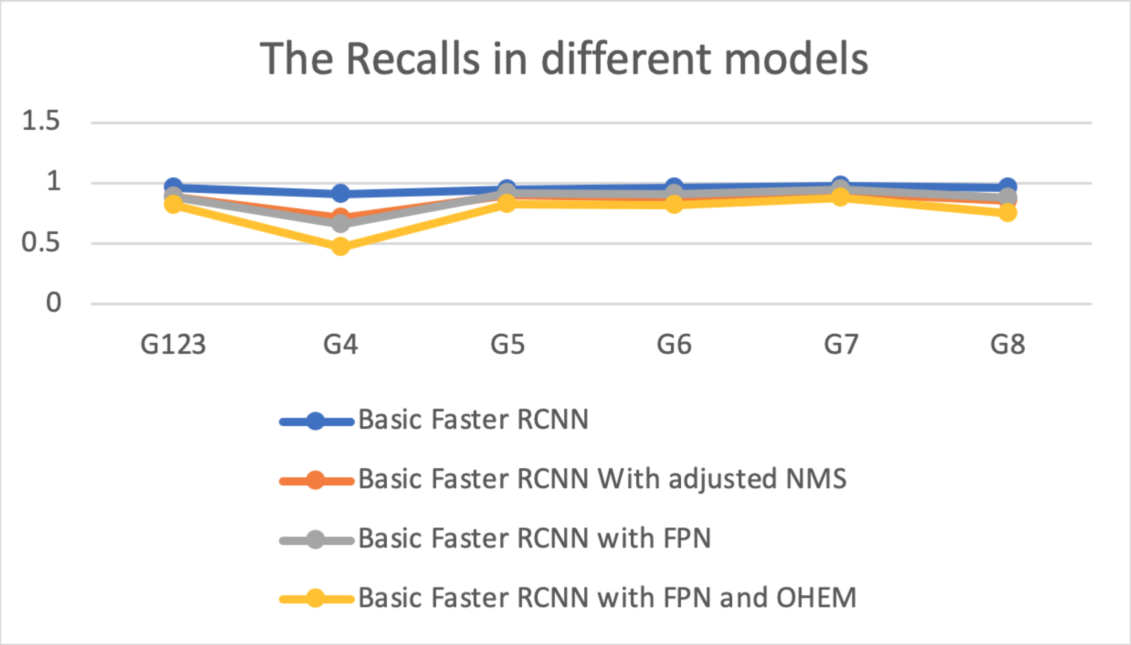


Figure 3.12: The change of Recall value with model structures modified

As a summary of the above investigation, Fig. 3.12 and Fig. 3.13 show the changes of recall and precision value when the microatoll detection model is modified. For the result in Fig. 3.12, although the recall decreases significantly for area G4, the recall for other areas is well maintained. While in Fig. 3.13, it can be seen that the precision is significantly improved due to the adjusted NMS and the use of FPN and OHEM.

Finally, another experiment also investigates the generalisation capability of the model in detecting microatolls from the areas of another island. The Three Isles island photographs taken by the same drone and labelled by human experts are used as the test set to assess the generalisation ability of the detection model trained above. The Three Isles island is geographically close to Nymph Island (as shown in Fig. 3.15). They are comparable to the training set as they have approximately the same sea-level history and tide range. The main difference is the age of the reef flats. This difference will vary depending

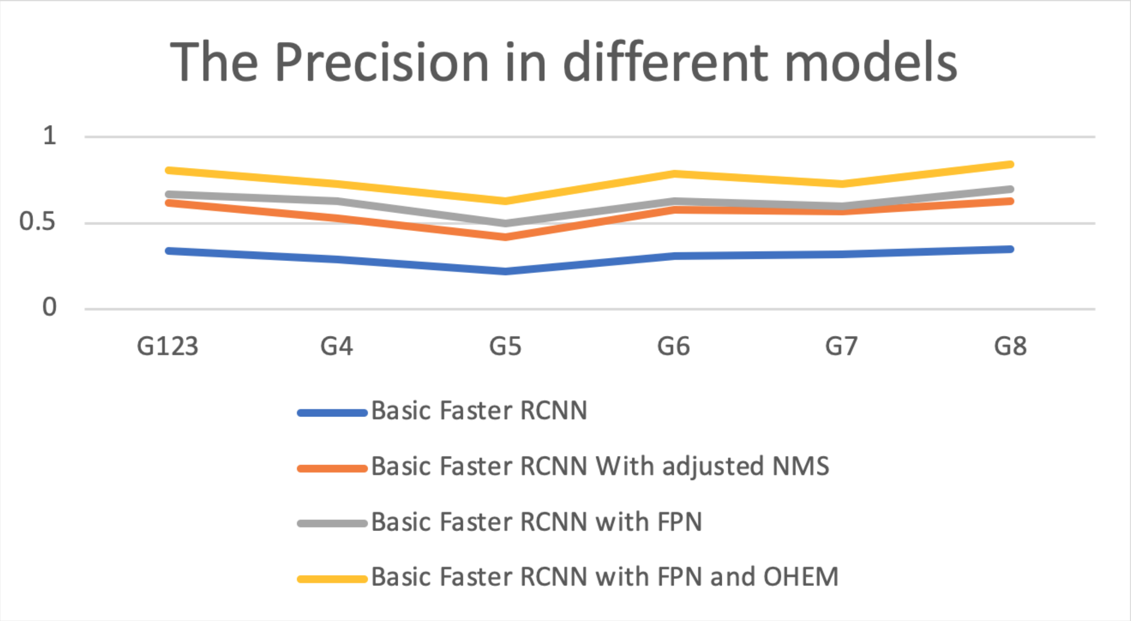


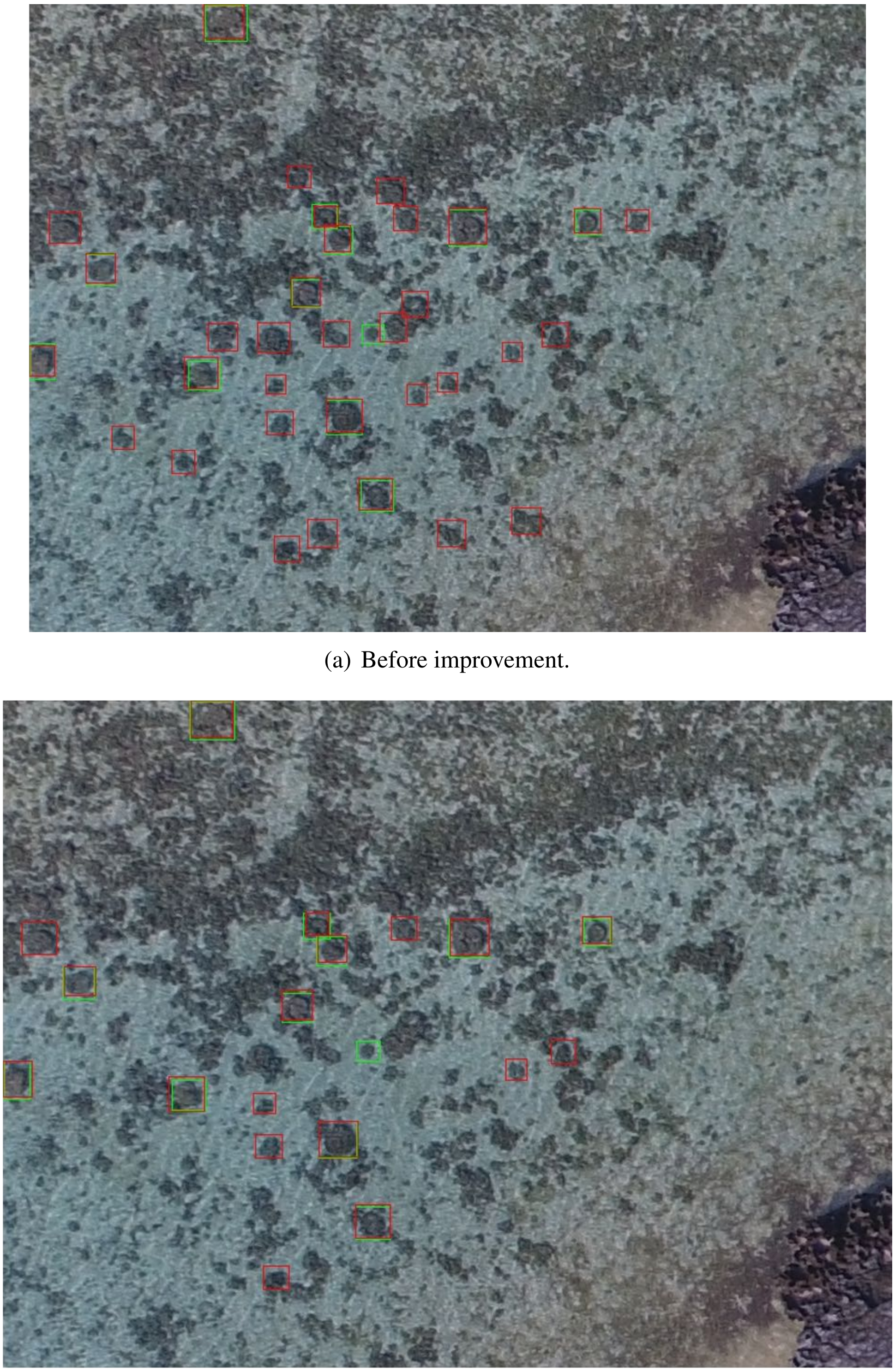
Figure 3.13: Performances enhancement in average (exclude G4)

on the age of the reef platform. These reef platforms have grown from the seafloor up to the current sea-level over the last 5,000-6,000 years (known as the Holocene). They have grown up from an underlying platform at a different height, which means that they have had a different length of time at sea-level. In terms of environmental processes, they are very similar. The detection results are presented in Table 3.7 and Fig. 3.16.

Model	Ground Truth	Detected	True Positive	False Positive	Recall	Precision	Average Precision
Basic_Faster_R-CNN With FPN and OHEM	3111	5098	2923	2175	0.94	0.57	0.72

Table 3.7: The performance of the above object detection model on the Three Isles island.

Fig 3.16 shows the detection result given by the above object detection model on the Three Isles island. The red and green boxes appear in pairs in most cases, which means most microatolls are correctly detected. However, the overall detection performance shown in Table 3.7 is not as promising as what is seen in Fig. 3.16. After examining the results, it is found that the object detection model works well for the areas that have microatolls. Some regions occupied by the mangroves are wrongly detected as microatolls because they display similar patterns as microatolls in the images captured by the drone. Those 'novels' areas (i.e. they are not similar enough to the set of images used to train the above detection model) affect the overall detection performance. A straightfor-



(a) Before improvement.

(b) After improvement.

Figure 3.14: The comparison of the detection results for the same area for different settings investigated above. The top is the result of basic setting (shown in Table 3.4) while the bottom is the result from the adjusted model (shown in Table 3.5).

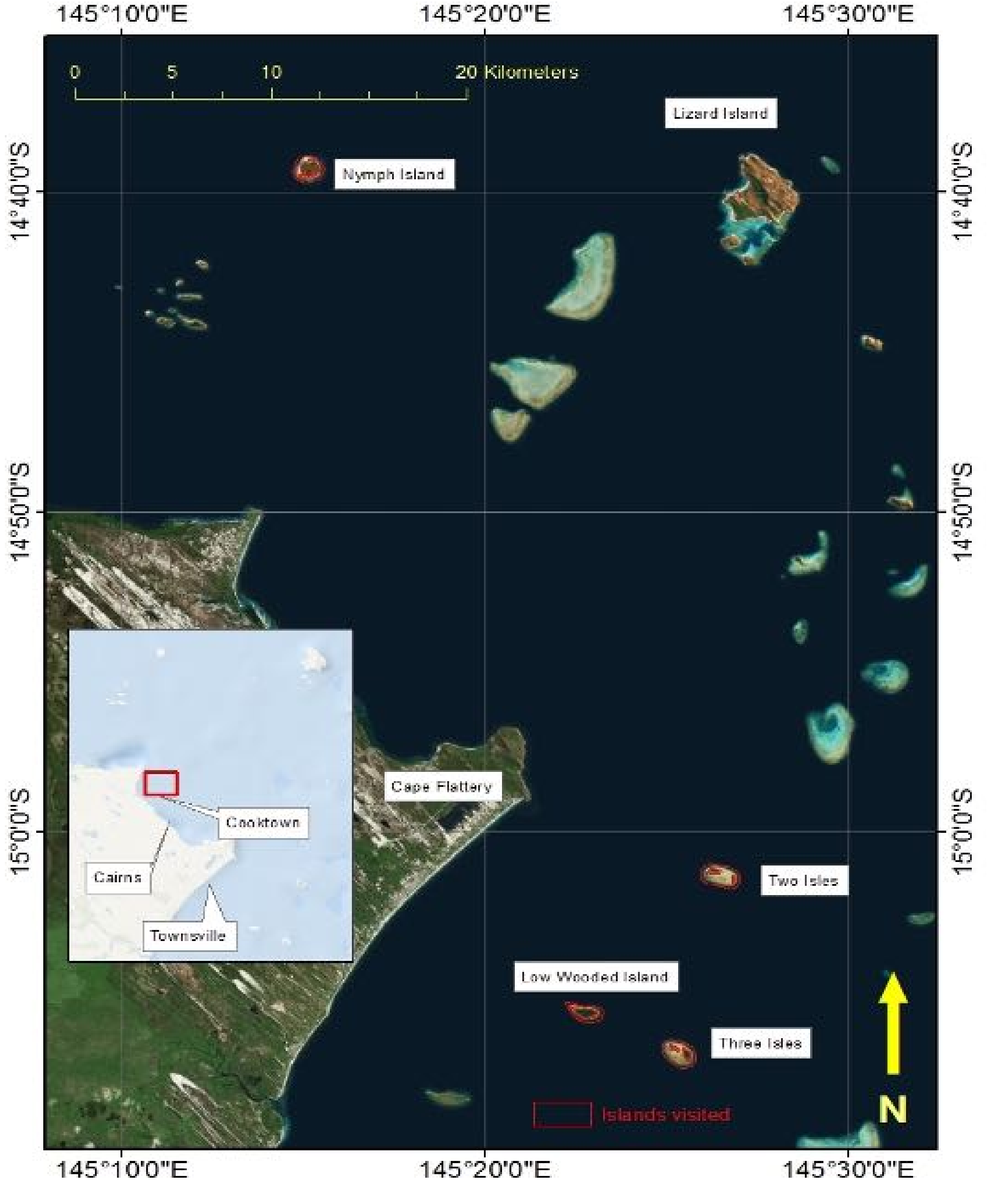


Figure 3.15: This map shows that Nymph island and Three Isles island have similar latitude and longitude coordinates.

ward solution may include more images of mangrove-occupied regions into the training set to retrain the object detection model. Meanwhile, enlarging the images to include mangroves and other environments may promote the performances in both Region Proposal (RoI) and the following detection phases. Combined with sufficient training data, the model could be potentially applied to many (up to 2,400) other islands on the Great Barrier Reef and possibly other global reef islands.

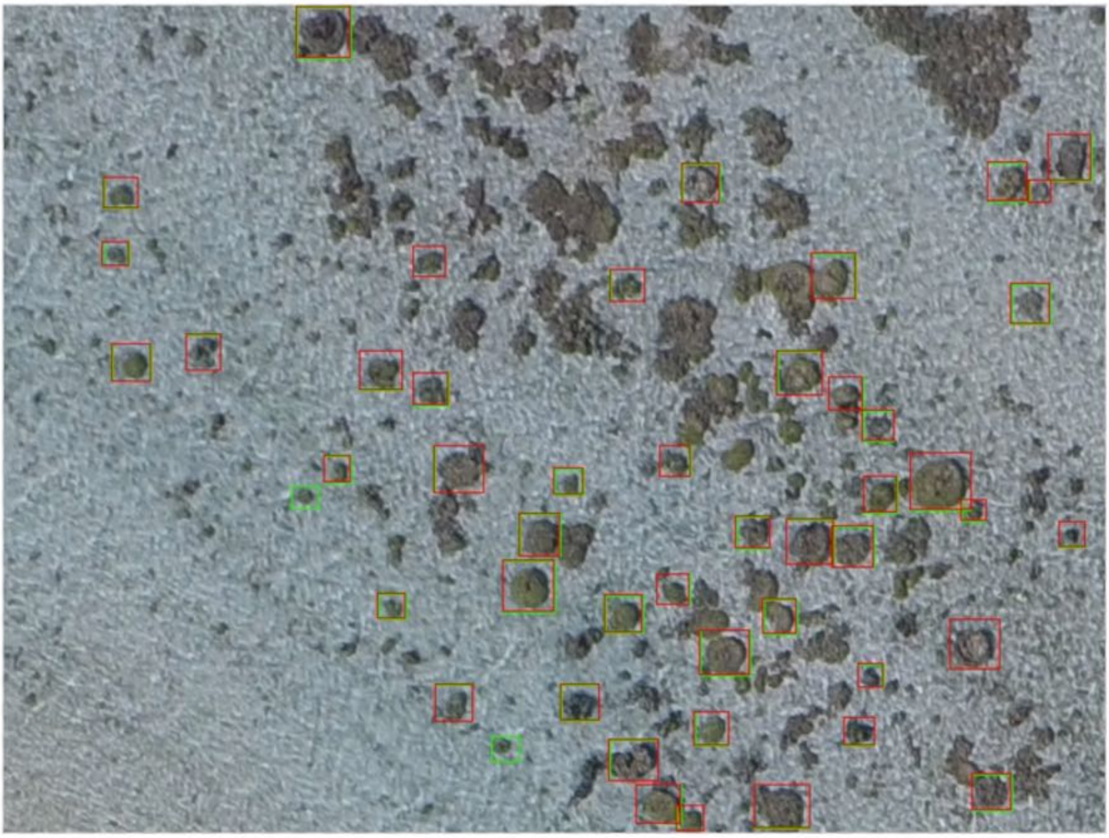


Figure 3.16: detection result in Three Isles island. Similarly, the green boxes are the ground truth and the red ones are detected results.

3.5 Proposed Method

3.5.1 Motivation

As the previous experimental study shows, the improvements from OHEM and adjusted NMS threshold take the cost of reducing recall and average precision. Meanwhile, those investigated models do not perform well in the most challenging area (G4). In sight of this, we examine and analyse the detection results in the most challenging area. The examination reveals that the critical factor of the performance degraded in this area is because the detection model is confused with individual microatolls and microatoll pavements. Comparing the individual microatolls and microatoll pavements, the significant difference between them is the background features. The individual microatolls are generally surrounded by water, which means the areas outside the individual microatolls have different features. Although the individual microatolls could have a similar colour to the background, their margins are distinguishable. In terms of the microatoll pavements, the areas outside the wrongly detected bounding boxes have the same unchanged features because individual microatolls connect to form microatoll pavements. In other words, the difference of surrounding features is the critical factor to distinguish individual microatolls from microatoll pavements. In the literature, many methods utilise context or background features to enhance object detection performances. Inspired by region average pooling [57], we found that the context features could achieve our aim of including the background information. These two factors prompted us to propose Microatoll-Net to use surrounding information to enhance the detection performance in the most challenging area.

3.5.2 Microatoll-Net

We propose Microatoll-Net that combines context information around individual microatolls to improve the detection from the most challenging area (G4).

The architecture of Microatoll-Net is illustrated in Fig. 3.17. In the literature, ResNet [34] strongly dominates the VGG [35] in the task of image recognition. Therefore, we take advantages of Faster R-CNN [11] but using ResNet [34] instead of VGG [35] as the backbone network for feature extraction. Besides, the FPN is also mounted on the Faster R-CNN. As aforementioned, to consider the background information, Kuan et al. [57] use RAP to include additional context features from all RoIs. Different from that work, a proposed module called context region of interest (C-RoI) actively extracts features from the area surrounding an RoI, which is generated by RPN in Faster R-CNN. This extra con-

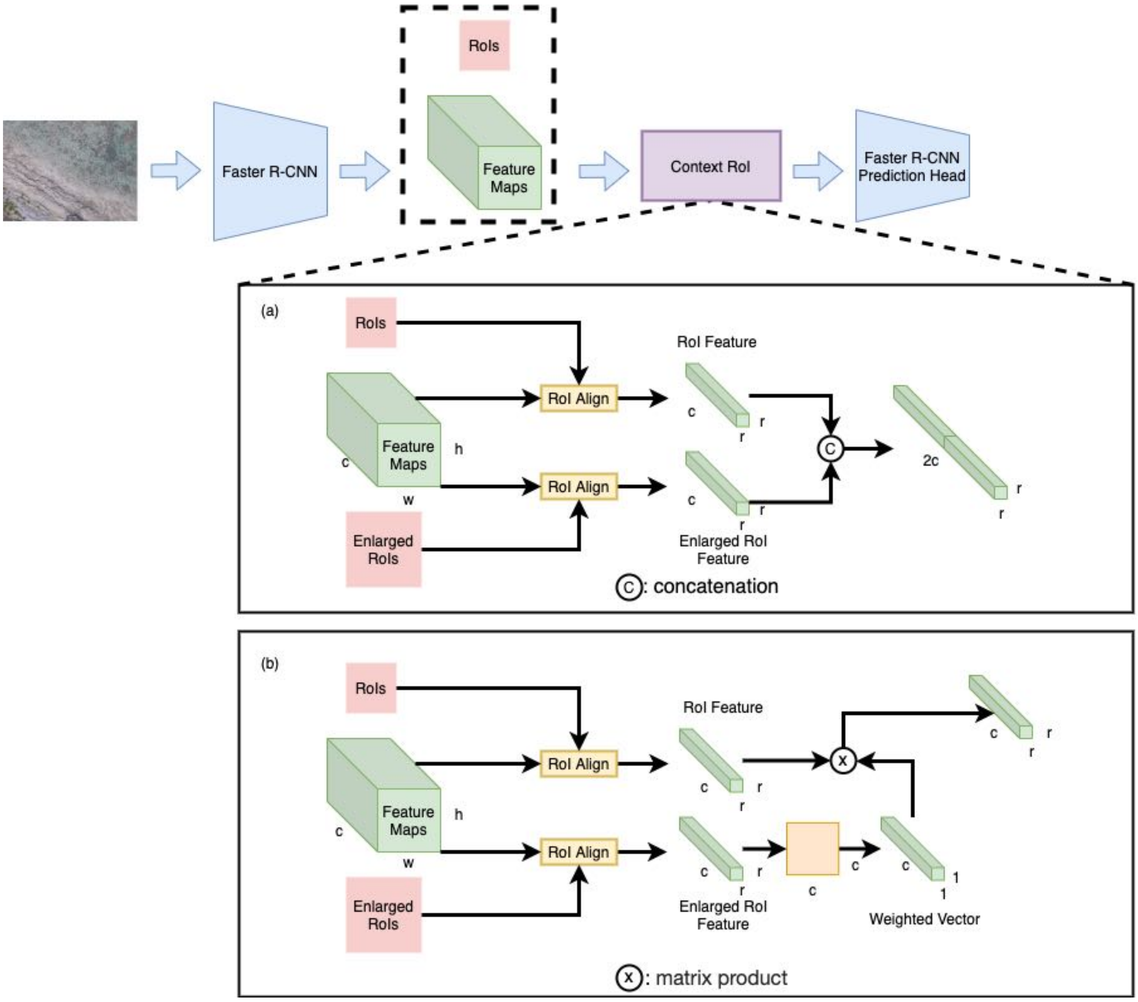


Figure 3.17: This figure illustrates the proposed Microatoll-Net. It takes an image as input and passes it through the Faster R-CNN module with FPN to generate a set of RoIs. They are then fed to the proposed C-RoI module (in purple) in which contextual features are extracted from each of the enlarged RoIs. The contextual features are integrated with those from the original RoIs and further fed to the detection head. Two integration ways are shown: (a) concatenating the two types of features; (b) using the contextual features as a signal via attention mechanism to weight original features.

textual information helps improve discrimination and boost detection performance.

As shown in Fig. 3.17, our C-RoI module expands the RoI Align [39] component in Faster R-CNN. The inputs of C-RoI are the RoIs with corresponding feature maps. Let the subscripts o and e denote “original” and “enlarged”, respectively. The RoIs are a set of n bounding boxes, each defined by its top-left and bottom-right coordinates (x, y, x', y'). Enlarging an RoI with respect to its centre so that the area of the enlarged RoI is $A_e = \alpha A_o$, where α is a preset parameter. Based on the same feature maps $\mathbf{X} \in \mathbb{R}^{c\times h\times w}$, cropping each RoI and its corresponding enlarged RoI, respectively. Passing them through the RoI

Align component produces their feature maps (of the same size) as $\hat{\mathbf{X}}_o, \hat{\mathbf{X}}_e \in \mathbb{R}^{c \times r \times r}$ where c is the number of feature channels and r is the height and width. Now, the issue boils down to how to integrate them. Four different methods are proposed as follows.

Obtaining the microatoll features $\hat{\mathbf{X}}_o$ and $\hat{\mathbf{X}}_e$ leads to a natural thought of how to efficiently utilise them. The first method is a simple concatenation, as shown in the subfigure (a) in Fig. 3.17. Considering $\hat{\mathbf{X}}_o$ and $\hat{\mathbf{X}}_e$ are in the same shape, and they are concatenated along the first dimension to obtain a feature map $\mathbf{X}_c \in \mathbb{R}^{2c \times r \times r}$, which is then fed to the detection head. Note that, original Faster R-CNN takes a feature map in the shape of $c \times r \times r$ as input. Hence, we add a convolutional layer (by having 1×1 kernel and setting stride as 1) to downsample the input feature maps for the original Faster R-CNN detection head. This model is named "M-Net" in short.

The second method is shown in the subfigure (b). It is motivated by the recent work on second-order pooling (SOP) and attention mechanism. Inspired by Gao et al. [59], in which SOP is utilised to generate an attention signal from the contextual feature $\hat{\mathbf{X}}_e$ to weight the features of the corresponding original RoI so as to obtain context-guided features for detection. Specifically, let $\mathbf{X}_r \in \mathbb{R}^{c \times d}$ denote a reshaped $\hat{\mathbf{X}}_e$, where $d = r \times r$. To obtain the SOP feature, the first step is computing an outer product $\mathbf{\Sigma} = \mathbf{X}_r \mathbf{X}_r^\top \in \mathbb{R}^{c \times c}$, which essentially characterises the correlation among each pair of feature channels. After that, a row-wise convolution Conv$(\cdot)$ is applied to the matrix $\mathbf{\Sigma}$ to obtain the attention vector $\mathbf{V} \in \mathbb{R}^{c \times 1 \times 1}$. The context-guided feature map is then calculated as $\mathbf{V} \otimes \hat{\mathbf{X}}_o$, where $\otimes$ means the components of $\mathbf{V}$ are multiplied to each channel. This model is named "M-Net1".

Additionally, another two ways are explored to investigate the pure efficacy of SOP features in Microatoll-Net, without considering the contextual information. One is similar to the approach of Gao et al. [59], only the original RoI features $\hat{\mathbf{X}}_o$ are used to generate SOP and the attention signal to weight itself. The variant is named "M-Net2". The other one directly feeds the SOP features obtained based on $\hat{\mathbf{X}}_o$ into the detection head. The variant is named "M-Net3". All of these variants are investigated in experimental study.

3.6 Experimental Result

Model Group	Baseline	Baseline+C-RoI	Baseline+RAP [57]	Baseline+FPN	M-Net1	M-Net2	M-Net3	M-Net
G123	0.80	0.80	0.80	0.88	0.89	0.89	0.88	**0.89**
G4	0.53	0.55	0.53	0.62	0.65	0.65	0.64	**0.66**
G5	0.71	0.72	0.72	0.80	0.81	0.81	0.79	**0.81**
G6	0.75	0.76	0.75	0.85	0.88	0.87	0.86	**0.87**
G7	0.8	0.81	0.81	0.89	0.89	0.90	0.89	**0.90**
G8	0.68	0.67	0.68	0.86	0.87	0.87	0.86	**0.87**
Average	0.712	0.718	0.715	0.817	0.832	0.832	0.820	**0.833**
FPS	36.3	33.4	34.2	31.4	7.7	7.2	6.5	**31.3**

Table 3.8: This table reports Average Precision (AP) of each detection method tested on each group of our microatoll dataset. “Baseline” is Faster R-CNN, “Baseline+C-RoI” is Baseline+our C-RoI module, “Baseline+RAP” is Baseline+region average pooling in [57], “Baseline+FPN” is Baseline+feature pyramid network. “M-Net” and “M-Net1, 2, 3” are the four variants of our Microatoll-Net defined in Sec. 3.5. The last row presents the detection speed of these detection methods in terms of the number of processed frames (i.e. test samples) per second.

3.6.1 Experimental Setting

With our microatoll dataset, object detection algorithms are tested in the “leave-one-out” setting [61]. Each of the eight groups is used as the test set in turn while the remaining groups are used for training.

The proposed Microatoll-Net is trained on two NVIDIA 1080Ti GPU cards for 12 epochs, with ImageNet pre-trained ResNet50. Each mini-batch has two samples per GPU card. Weight decay is 0.0001, momentum is 0.9, and a learning rate warm-up strategy [34] is used. α is set as 2.0 to enlarge an RoI in our C-RoI module.

3.6.2 Experimental Analysis

Table 3.8 shows the comparison of eight models, explained in table caption, in terms of the average precision (AP) obtained on each test group. The left part shows the effect of adding the C-RoI, RAP or FPN module to the baseline Faster R-CNN. As seen, adding our C-RoI module has boosted detection, particularly on the most challenging group G4 (0.53 to 0.55). Also, C-RoI outperforms RAP on G4 and the averaged AP, showing its

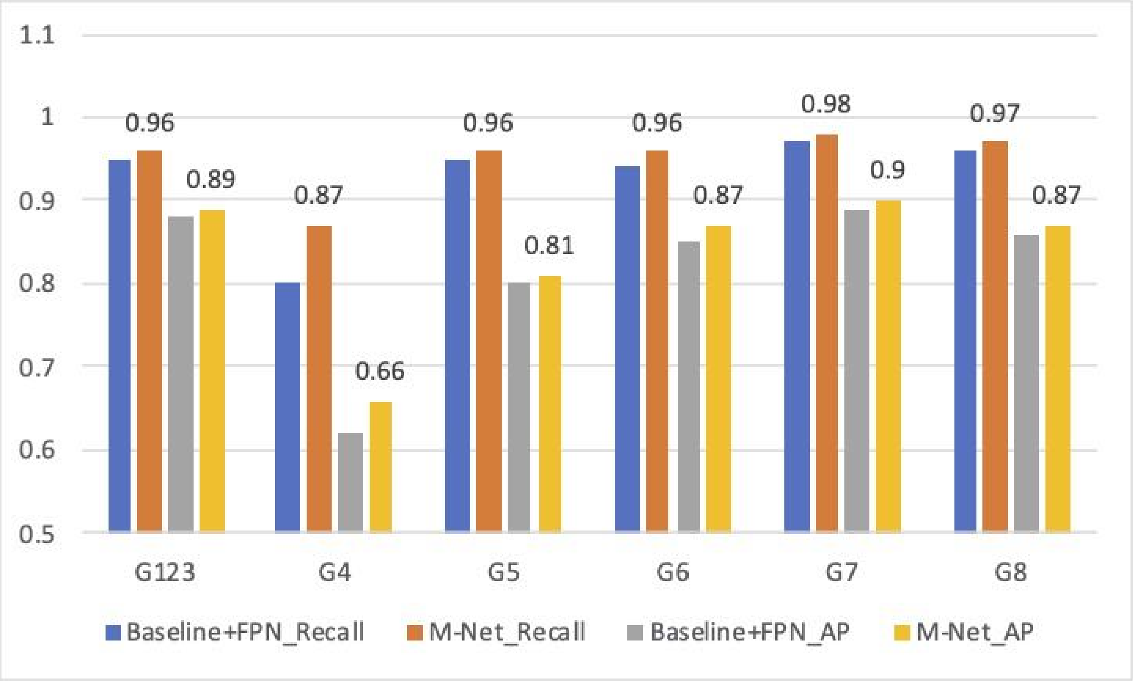


Figure 3.18: The Average Precision (AP) and Recall of Baseline+FPN and Microatoll-Net on each test group. The largest improvement is obtained on group G4. Best viewed in color.

suitability for microatoll detection. Meanwhile, adding FPN achieves the most considerable improvement, confirming its importance in microatoll detection. The right part further verifies the efficacy of our C-RoI module, even when FPN is used. As shown, all four variants of M-Net outperform baseline+FPN. Particularly, the most significant gain is obtained on G4 (0.62 to 0.66), which aligns well with our motivation to design the C-RoI module.

Comparison of the four variant of M-Net shows that they achieve overall similar performance. This indicates that all variant models help microatoll detection. This indicates that simple concatenation, attention with the SOP on the contextual features, and even the SOP on the original RoI features all help microatoll detection. This suggests that contextual information and SOP features could be complementary in enhancing discrimination and that their deep integration could be pursued in the future work. Finally, taking the detection speed shown in the last row into account, M-Net in the last column becomes a clear winner in terms of accuracy and efficiency. Also, it well maintains the detection speed when compared with baseline+FPN. To provide more details, it also reports that the average precision and recall values of two typical models among the compared ones in Table 3.8, baseline+FPN and M-Net, on each test group. As shown in Fig. 3.18, our M-Net consistently attains higher AP and Recall on all groups, especially on the most challenging group G4.

3.7 Chapter Summary

This chapter reports our recent work on utilising visual object detection to locate microatolls from drone images, and it is significant for coastal and climate research. A solid benchmark dataset is created for microatoll detection from drone images. It investigates existing modules and proposes a Microatoll-Net that better considers the characteristic of microatolls. The experimental study demonstrates the promising performance of our method and the more work needed to address the detection in challenging areas. Future works could focus on enhancing the performance of the problematic areas and utilising the model on other Great Barrier Reef islands.

Chapter 4

Microatoll-GAN for Microatoll Detection

Contents

4.1 Introduction

The previous chapter presents the experimental study on our microatoll dataset to investigate the performance of Faster R-CNN [11] and its variants (i.e. combining FPN and

OHEM) on the task of microatoll detection from drone images. After that, Microatoll-Net is proposed that adds a novel microatoll context RoI module upon the Faster R-CNN to fuse the local microatoll feature and corresponding context information around microatolls. The proposed Microatoll-Net surpasses the fine-tuned Faster R-CNN in the most challenging area (G4). Microatoll-Net and Faster R-CNN variants are region-based methods that generate a set of RoIs to reduce the object detection task to an image recognition problem. Meanwhile, the anchors in such detectors are translation invariant, and the anchors minimise the parameters in the region proposal phase to reduce the overfitting risk. The success of anchors leads them to be extensively used in one-stage detectors. However, there are two drawbacks to the use of anchors in object detectors. First, a large number of anchors are generally needed to sufficiently cover the ground truth bounding boxes of objects. Second, the scales and aspect ratios of the anchors are hyper-parameters that rely on manual design. In addition to the anchors, non-maximum suppression (NMS) is required to remove those overlapping boxes and ensure one object will be detected once after the detection phase—the selected threshold for non-maximum suppression severely impacts the detection performance in real-world applications.

This chapter reports our recent research on exploring the approach of centre-based microatoll detection from drone images without anchors or post-processing steps (i.e. NMS). As described in the literature, a microatoll is a circular coral colony that can live for thousands of years [2]. The circular shape arouses great interest in presenting microatolls by their centres instead of bounding boxes. Kong et al. proposed FoveaBox [30] that transforms the bounding boxes to smaller centre areas (they are named fovea areas) to eliminate the use of bounding boxes in object detection tasks. The fovea areas are accordingly computed from the ground truth bounding boxes, and their model achieves notable results compared with state-of-the-art methods. It inspires us to transform the bounding boxes into central areas which can reflect the information of object locations.

Recently, Wang et al. [13] proposed a conditional generative adversarial network based segmentation model which contains a Miss Detection vs. False Alarm loss function (named MDFA in short) to segment targets from infrared small object segmentation (ISOS) datasets. The targets in the ISOS datasets only occupy a few pixels in grey scale images, and the MDFA model successfully segments them by combining the local and global features. Since the data patterns of small targets can be learned through this method, identifying whether this method can learn a more complex data pattern such as microatoll centres becomes an interesting research problem. The idea of transforming bounding boxes to object centre density maps comes from CenterNet [29]. It uses a Gaussian kernel to obtain two-dimensional Gaussian distribution centre density maps as regression targets. Following the above idea, a generative adversarial network based framework to detect microatoll centres is proposed, named Microatoll-GAN. With the same structure of MDFA, Microatoll-GAN has two generators and one discriminator. Each of the generators is designed to generate a centre density map from a given image, and they cooperatively reduce miss detection (by a generator with the smaller receptive field) and false alarm (by the other with a larger one). A discriminator network is to make the two density maps align with the ground truth detection result. Different to MDFA [13], in our case, annotations are first transformed from the bounding boxes in the microatoll dataset (created in Chapter 3) to centre-based density maps. Second, the receptive fields in both generators are modified according to the analysis of microatoll sizes. Specifically, a generator uses a smaller receptive field to utilise the local feature to reduce miss detection (MD), and the other has a bigger one to use the surrounding feature to reduce false alarm (FA). The discriminator is changed to downsample the input images, and the max-pooling layers are removed to weaken the ability of the discriminator. Weakening the ability of the discriminator means that an overstrong discriminator will cause the training generator to fail because of the vanishing gradients. Hence, we adaptively modify the discriminator to ensure a proper training procedure. The proposed Microatoll-GAN increases recall from 25% to 55% and precision from 57% to 64% compared with the original MDFA.

In sum, a novel generative adversarial network based microatoll detection network is proposed that successfully converts a segmentation model to serve the microatoll detection task. The model significantly elevates the performance of MDFA in the microatoll centre detection task and removes the use of anchor and post-processing step in

Microatoll-Net.

4.2 Related Work

Wang et al. [13] describe the concepts of miss detection (MD) and false alarm (FA) in infrared small object segmentation. This idea grabs great interest since we aim at detecting microatoll centres. In the scenario of segmenting objects from ISOS datasets, the targets are pixel-level. Therefore, the MD indicates that those foreground pixels are not included in segmentation results, while the FA describes the background pixels are improperly labelled as foreground. MD and FA are formally defined as follows:

$$\begin{aligned} MD &= \| (\mathbf{S}_1 - \mathbf{S}_{gt}) \otimes \mathbf{S}_{gt} \|_2^2 \\ FA &= \| (\mathbf{S}_2 - \mathbf{S}_{gt}) \otimes (1 - \mathbf{S}_{gt}) \|_2^2, \end{aligned} \tag{4.1}$$

where $\mathbf{S}_1, \mathbf{S}_2$ and $\mathbf{S}_{gt} \in [0,1]^{w \times h}$ are segmentation results respectively from the generator with smaller receptive field G_1 and the other with lager one G_2. Additionally, $\mathbf{S}$ is used to broadly represent any arbitrary $\mathbf{S}_1, \mathbf{S}_2$ or $\mathbf{S}_{gt}$. The ground truth heatmaps are in the same size of $w \times h$ with input images, the L_2 norm is matrix-based, the operation $\otimes$ denotes element-wise multiplication and the operation — means subtraction. Based on the MD and FA definitions, they construct a model that consists of two generators and a discriminator. A generator $G_1(\cdot)$ takes an input image $\mathbf{I} \in \mathbb{R}^{3 \times w \times h}$ to generate a heatmaps $\mathbf{S}_1$ by $G_1(\mathbf{I})$. Similarly, $G_2(\mathbf{I})$ indicates the result $\mathbf{S}_2$ from the other generator. The discriminator ($D(\cdot)$) takes $\mathbf{I}$ and the corresponding heatmaps $\mathbf{S}_{gt}, \mathbf{S}_1$ and $\mathbf{S}_2$ to decide if they come from G_1, G_2 or the ground truth.

$$Loss_{(D,G_1,G_2)} = \min(\alpha_1 L_{GC} + \alpha_2 (L_{MF_1} + L_{MF_2}) + \alpha_3 L_{cGAN}) \tag{4.2}$$

As shown in Equation 4.2, the loss function used in MDFA has four terms, and each of them serves a specific purpose. Firstly, L_{GC} is to minimize the difference between $\mathbf{S}_1$ and $\mathbf{S}_2$. $\mathbf{I}$ and the corresponding heatmap $\mathbf{S}$ are sent to D, and then D outputs $D((\mathbf{I}, \mathbf{S})) \in \mathbb{R}^{1 \times 3}$

that presents the source of $\mathbf{S}$ in a binary format. Recall that $\mathbf{S}_{gt}, \mathbf{S}_1$ and $\mathbf{S}_2$ denote the ground truth, $G_1(\mathbf{I})$ and $G_2(\mathbf{I})$ heatmaps, respectively. The L_{GC} is defined as

$$L_{GC} = ||\phi(D((\mathbf{I}, \mathbf{S}_1))) - \phi(D((\mathbf{I}, \mathbf{S}_2))||^2, \tag{4.3}$$

where each of the terms presents a discriminator ($D(\cdot)$) that takes an concatenation of an image and corresponding a heatmap as input to generate a feature map ϕ for minimising the difference of feature maps between G_1 and G_2.

Secondly, L_{MF} originally aims to narrow the difference between MD and FA to make two generators to generate the same heatmaps. Let L_{MF_1} and L_{MF_2} denote the L_{MF} for G_1 and G_2, respectively. In the literature, G_1 is to reduce the MD by using a smaller receptive field and G_2 is to minimise FA by using a larger one. Thus, L_{MF_1} and L_{MF_2} are in the same form, but they have different focuses. The difference is reflected in the formula through the λs. As shown in Equation 4.4, the aims are achieved by using λ_1 and λ_2 to weight MD and FA. Although L_{MF_1} and L_{MF_2} are originally designed to reduce MD and FA, the information exchange between G_1 and G_2 could, in turn, boost the ability of G_1 in reducing FA and similarly boost the ability of G_2 in reducing MD.

$$\begin{aligned} L_{MF_1} &= \frac{1}{n}\sum_{i=1}^{n}(\lambda_1 MD_{1i} + FA_{1i}) \\ L_{MF_2} &= \frac{1}{n}\sum_{i=1}^{n}(MD_{2i} + \lambda_2 FA_{2i}), \end{aligned} \tag{4.4}$$

where MD_{1i} and FA_{1i} represent the MD and FA coming from G_1, and this applies to MD_{2i} and FA_{2i} similarly.

Lastly, L_{cGAN} comes from conditional GAN [63] that originally combines its generated result with encoded text feature to fool a discriminator, while the discriminator behaves to identify if the given input is faked or not. Similarly, L_{cGAN} takes original images as conditions and adds the generated segmentation results to fool the discriminator. At the same time, the discriminator tries to distinguish the given inputs to decide if they are from the ground truth, G_1 or G_2. Formally, Equation 4.5 combines three binary cross-entropy losses (F_{BCE}) to achieve this goal:

$$L_{cGAN} = F_{BCE}(D((\mathbf{I},\mathbf{S}_{gt})),\hat{\mathbf{S}_{gt}}) + F_{BCE}(D((\mathbf{I},\mathbf{S}_1)),\hat{\mathbf{S}_1}) + F_{BCE}(D((\mathbf{I},\mathbf{S}_2)),\hat{\mathbf{S}_2}) \tag{4.5}$$

, where $\hat{\mathbf{S}_1}, \hat{\mathbf{S}_2}$ and $\hat{\mathbf{S}_{gt}}$ are a binary label to indicate the sources of $\mathbf{S}_1, \mathbf{S_2}$ and $\mathbf{S}_{gt}$, $F_{BCE} = -(y\log p + (1-y)\log(1-p))$, and y and p represent the ground truth and the prediction result, respectively.

Following the idea described above, a novel Microatoll-GAN is proposed for the microatoll centre detection. Because the targets in the ISOS datasets are at pixel-level, their proposed MD and FA are strict to the position of foreground pixels. Therefore, in the task of this chapter, following the original setting will lead the predicted centres to be considered as false positive results if they are not wholly and precisely the same as the corresponding ground truth targets. For this reason, instead of using the pixels as targets, the bounding box annotations in our microatoll dataset are transformed to centre density maps mentioned in [29]. We use this transformation for microatoll detection due to the fact that as long as the detected centres by the model are within the limits of bounding box, it can indicate the location of the microatoll in the Coral reefs of oceans. It is not required for the centres detected by the model to be the exact centres of the of the microatoll blobs. An experimental study is conducted to examine the performance of MDFA in two different annotation formats to draw forth the necessary annotation transformation.

4.3 Proposed Method

4.3.1 Motivation and Challenge

The Microatoll-GAN is motivated by MDFA [13], a segmentation model designed to segment pixel targets from infrared images via two different models to reduce the miss detection and the false alarm, respectively. In short, the MDFA model generates two heatmaps from two generators and align them with the ground truth by a discriminator. The two generators actively extract local and global features to achieve this aim. It is interesting in employing the same structure to detect microatoll centres because individual microatolls are circular, and their centres could well indicate their locations.

However, there are three challenges for adapting this segmentation model to be a detection one. First, the samples in ISOS and our microatoll datasets are significantly different. As mentioned in the work of Wang et al. [13], the ground truth $\mathbf{S}_{gt} \in [0,1]^{w\times h}$ indicates that the pixels are one if they are foreground. In contrast, those background pixels are set to zero. While in our microatoll dataset, the analysis in Fig. 3.8 and Fig. 3.9 shows the heights and widths in most microatoll annotations laying in the range of 25 to 75 pixels. In our experimental study, it is found that using a few pixels to represent a centre or setting all pixels within a bounding box as one will lead to detection failure (see Fig. 4.1). Due to these, the bounding box should be better transformed into a central area with the highest value to present the mathematical centre of a bounding box, and this value gradually decreases to 0 when the pixels are away from that centre. Inspired by CenterNet [64], a Gaussian kernel $Y_{xyz} = exp(-\frac{(x-\hat{p})^2+(y-\hat{p})^2}{2\sigma_p^2}) \in [0,1]^{w\times h}$ is applied to represent the density map in size of $w \times h$ (let $w,h = R$ because the w and h are the same for input images), where $\hat{p}$ is a four times low-resolution equivalent ground truth centre point and σ_p is an object size-adaptive standard deviation [27].

Secondly, the difference between targets in ISOS and our microatoll datasets leads to the modification of the model receptive fields. As mentioned in the literature, the receptive field is the area where a filter learns from an input image. In MDFA, the two sizes are 31×31 and 257×257 in G_1 and G_2, respectively. Recall the targets in ISOS datasets are of pixel-level and shown in the grey scale image, so the target features are limited. Hence, the authors employ a dilated convolutional layer [65] to extract context feature to enrich the pixel-level target features in G_1 for reducing MD. Because targets are sparse (i.e. one image only contains one target), the receptive field in G_2 covers an entire image to use context features to reduce FA. While in our microatoll dataset, each of the microatolls is presented in a high-resolution drone image. The features of microatolls are sufficient for reducing MD. Thus, the use of a dilated convolutional layer is not necessary, and a receptive field close to most microatoll sizes is sufficient to extract

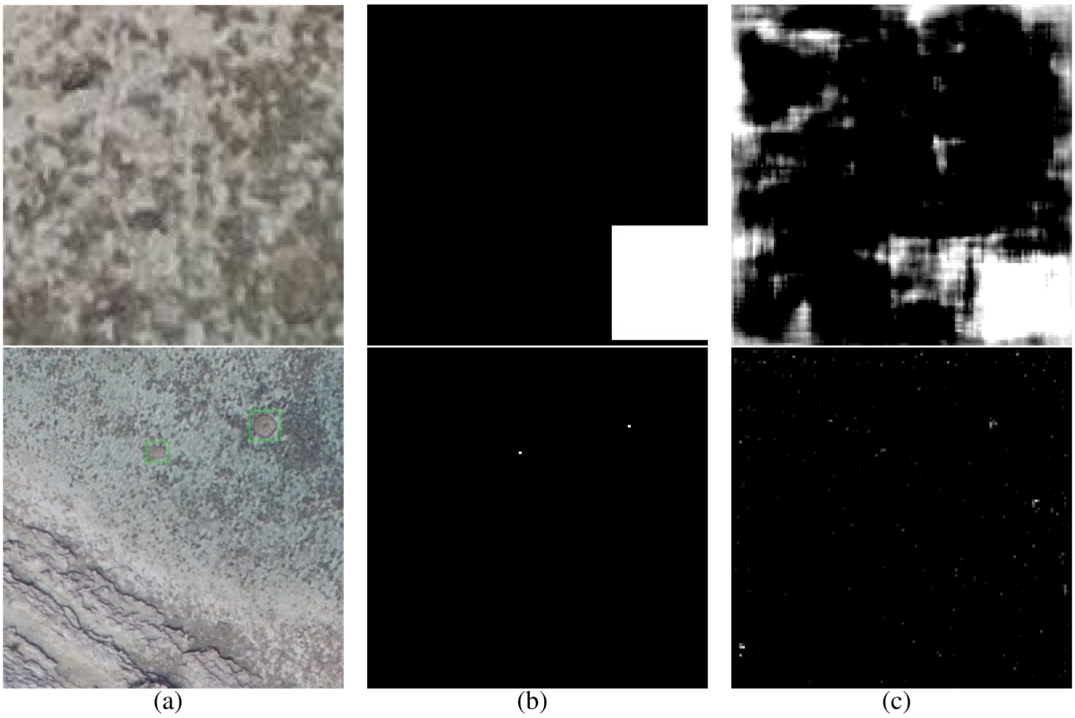


Figure 4.1: This figure indicates the detection failures in transforming a bounding box to a single centre pixel or regarding all pixels within a bounding box as a detection target. Column (a) to (c) represent image, regression ground truth and prediction results, respectively. The first row illustrates that making every pixel within the bounding box as one causes detection failure. Also, the other failure is shown in the second row, which uses a pixel to present the object centre.

local microatoll features. In terms of G_2, the original receptive field will bring additional noise. The noise is caused by extracting external features from an individual microatoll. The targets in ISOS datasets are sparse, and most images only contain one target. As for our microatoll dataset, most images contain more than one target. That means using a larger receptive field like the original G_2 to cover the entire image will lead filters to learn noise patterns from the background and other microatolls when extracting a feature from an individual microatoll. Therefore, reducing the size of the receptive field in G_2 to meet the requirements of enclosing surrounding information and avoiding additional noise is challenging and important. In sum, based on the analysis shown in Fig. 3.8, the receptive fields in G_1 and G_2 are enlarged to 50×50 and reduced to 175×175, respectively.

The last challenge is how to evaluate the detection performance properly. For the infrared images in ISOS datasets [13], the segmentation result is presented in a mask with the same size as an input image. Because the targets are of a few pixels and have exact locations in the image, using the absolute difference of foreground pixels could efficiently compute the true positive and false positive. While in microatoll centre detection, the detector is required to output the locations of microatolls by giving indicators around the microatoll centres instead of exactly locating the microatoll centres. This difference needs us to design a suitable evaluation method to evaluate microatoll centre detection performance.

4.3.2 Proposed Microatoll-GAN

Data Processing

As the experimental study shown in Fig. 4.1, the modification of microatoll dataset (introduced in Chapter 3) follows the method used in CenterNet [64]. Instead of using 800×600 pixel patches, all the patches are cropped from the original drone images to have the size of 512×512 pixels. There are two reasons for selecting such a size: first, we follow the original MDFA model to take a square image as input. Moreover, sufficient surrounding information should be retained for microatoll detection. Second, the image in this size meets our hardware configuration to set a mini-batch as 2 per GPU.

The aforementioned Gaussian kernel is utilised to convert a bounding box to a centre density map during the annotation transformation process. As shown in Fig. 4.2, CenterNet [64] gives three cases of overlapping between the detected (in red) and ground truth (in green) boxes. To simplify the explanation, we take the last case in the figure as an example first. We let the threshold of intersection over union of red and green boxes to be $\beta = \frac{hw}{(h+2r_3)(w+2r_3)}$ where the h and w are the height and width of the red box and r_3 denotes the radius shown in the figure. Then we can rewrite the above equation as $4\beta r_3^2 + 2\beta(h+w)r_3 + (\beta-1)hw = 0$. Let $a = 4\beta, b = 2\beta(h+w)$ and $c = (\beta-1)hw$, and it can be obtained that $r_3 = \frac{-b+\sqrt{b^2-4ac}}{2a}$. That means a detected box, which is completely within the green box, will be considered to be correctly detected if the resulted radius in the figure $r \leq r_3$. Similarly, the second case, $\beta = \frac{(h-2r_2)(w-2r_2)}{hw}$, and we have $4r_2^2 - 2(h+w)r_2 + (1-\beta)hw = 0$. Let $a = 4, b = -2(h+w)$ and $c = (1-\beta)hw$, and we can obtain that $r_2 = \frac{-b+\sqrt{b^2-4ac}}{2a}$. That means a detected box, which is outside of the green box, will be considered to be correct if its radius $r \leq r_2$. In terms of r_1, $\beta = \frac{(h-r_1)(w-r_1)}{2hw-(h-r_1)(w-r_1)}$. Let $a = 1, b = -(h+w)$ and $c = \frac{(1-\beta)wh}{1+\beta}$, and we can then obtain that $r_1 = \frac{-b+\sqrt{b^2-4ac}}{2a}$. Therefore, a detected box which is partially overlapping with the green box will be considered to be correctly detected if the radius $r \leq r_1$. Finally, we select the smallest radius of r_1, r_2 and r_3 to create the center map of an object.

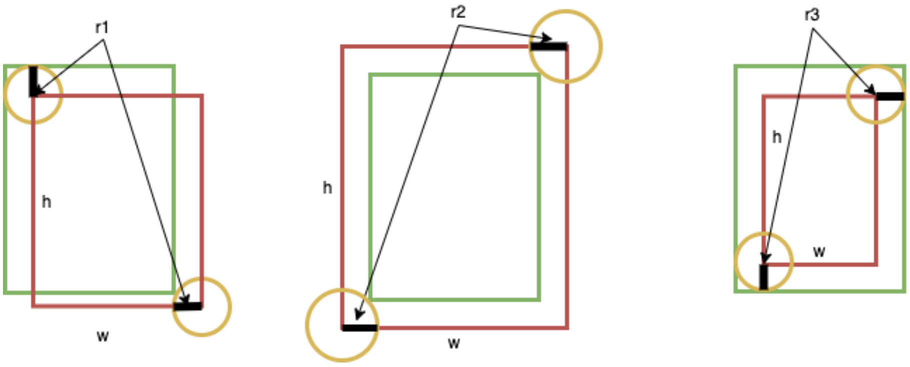


Figure 4.2: As introduced in CenterNet [64], there are three cases of the overlapping of the detected (in red) and the ground truth boxes (in green). This figure gives three critical values (three radii) for which the intersection over union of detected and ground truth bounding boxes are greater or equal to the threshold β: 1) the detected box is partially overlapping with the ground truth box with radius r_1; 2) the detected box is completely outside of the ground truth bounding box with radius r_2; 3) the detected box is completely within the ground truth bounding box with radius r_3. As shown in the figure, the h and w correspond to the height and width of predicted boxes while the r_1, r_2 and r_3 are radii for the three cases, respectively.

Finally, we transform the bounding boxes to corresponding density map represented by a two-dimensional Gaussian kernel as the ground truth. Note that the ground truth density maps are downsampled four times to avoid a similar detection failure because the centre areas occupy too many pixels (see example in the first row of Fig. 4.1).

Proposed Model

The proposed model consists of a discriminator (D) and two generators (G_1 and G_2) as the model described in MDFA [13]. Each of the generators maps an input image $\mathbf{I}$ to a density map $\mathbf{S}$ that shows the locations of microatolls, subject to the minimization of MD or FA. At the same time, the D classifies the sources of those density maps by taking the corresponding input image as the condition, aiming to align the two density maps

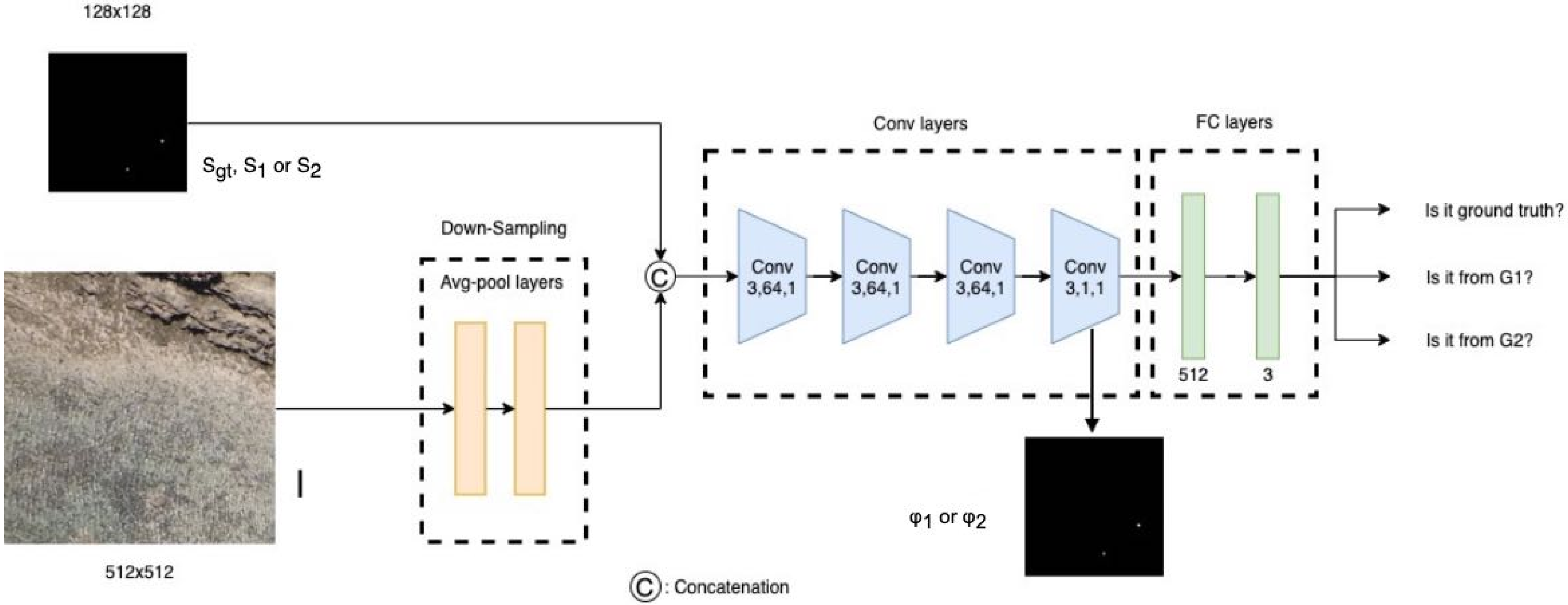


Figure 4.3: The overview of Discriminator (D). The D takes 512×512 images and a 128×128 center density maps ($\mathbf{S}_{gt}$, $\mathbf{S}_1$ or $\mathbf{S}_2$) as input. Images are firstly downsampled to the same size with centre density map, then they are concatenated (the C operation) for the following sub net, which is to tell where the center density map comes from. Meanwhile, the last convolutional layer will output the ϕ_1 and ϕ_2 to minimise the L_{GC}.

to the ground truth. The architecture of D is illustrated in Fig. 4.3. The D takes $\mathbf{I}$ and $\mathbf{S}$ to decide the $\mathbf{S}$ from G_1, G_2 or the ground truth. In the original MDFA, the D consists of two max-pooling layers, four convolutional building blocks and three fully connected layers in order. A similar structure is used in our case but modified to meet the following requirements: i) weakening the discriminator to avoid gradient vanishing of generators and ii) aligning the sizes of $\mathbf{I}$ and $\mathbf{S}$ for concatenation. According to Arjovsky and Bottou [66], a stronger discriminator may lead to gradient vanishing of a generator. Therefore, the two max-pooling layers are replaced with two average-pooling layers so that the sharp features may not be identified, while max-pooling may help build a stronger discriminator by sharping the features. Different to MDFA that $\mathbf{I}$ and $\mathbf{S}$ have the same width and height, in our model, the $\mathbf{I}$ is four times bigger than $\mathbf{S}$ after previously described data processing in our microatoll dataset. Hence, the two average-pooling layers are applied to downsample the input image only. The downsampled $\mathbf{I}$ is concatenated with $\mathbf{S}$ and the result is sent to four convolutional building blocks in order (see the blue block in Fig. 4.3). Besides, the D also corresponds to minimising the gap between G_1 and G_2. Specifically, in Fig. 4.3, the $128\times$ 128 density map $\mathbf{S}$ could represent $\mathbf{S_{gt}}$, $\mathbf{S}_1$ and $\mathbf{S}_2$. ($\mathbf{I}$, $\mathbf{S}_{gt}$) , ($\mathbf{I}$, $\mathbf{S}_1$) or ($\mathbf{I}$, $\mathbf{S}_2$) are sent to D to obtain feature maps ϕ (ϕ_1 or ϕ_2) for $\mathbf{S}$ from the last convolutional building block to compute the L_{GC} value for minimising the gap of density maps from G_1 and G_2. Each of the convolutional building blocks is serially made up of a convolutional layer, a batch normalisation and an activation function (leaky-ReLU [67]). All convolutional building blocks in this chapter have the same structure unless otherwise stated.

As the bounding box analysis shown in Section 3.8, most of the microatolls are of 50 pixels in height and width. Meanwhile, Section 4.3.1 shows that the receptive fields of G_1 and G_2 should be adjusted to 50×50 and 175×175, respectively. The difference in receptive field sizes indicates that G_1 focuses on extracting local features of microatolls to reduce MD, and the other is used for the surrounding features to reduce FA. Additionally, the MDFA uses dilated convolutional layer to enlarge the receptive field to enclose context feature to enrich the features of pixel-level targets. However, in our microatoll dataset,

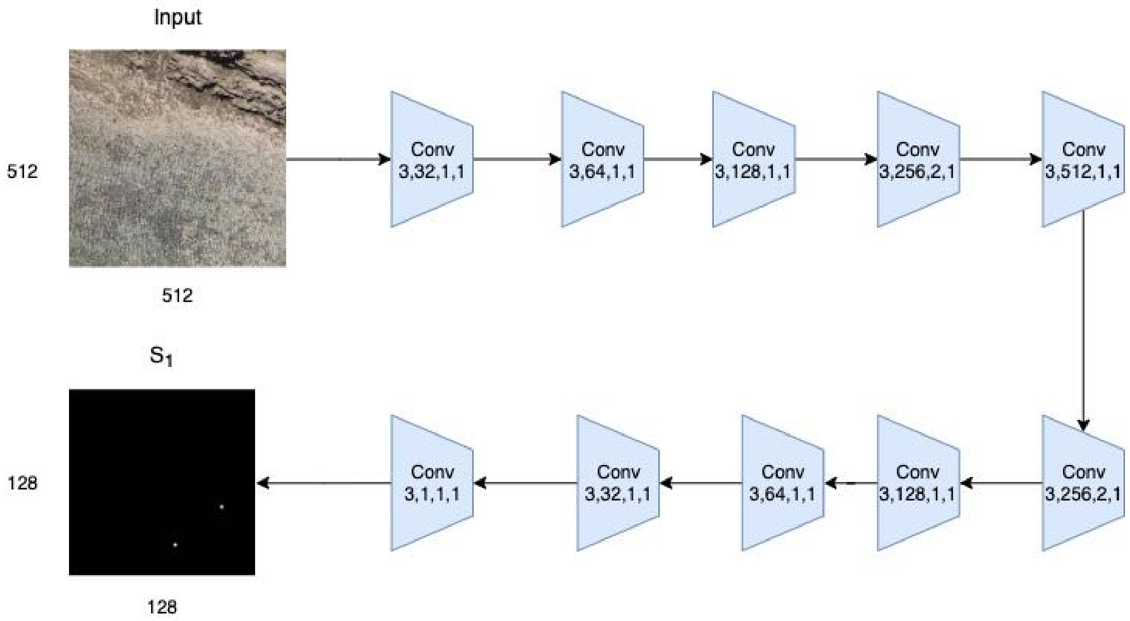


Figure 4.4: The structure of the generator which is used to reduce miss detection, named G_1. Each of the numbers in the block represents kernel size, channel, stride and dilation, respectively. Feeding a 512×512 image to G_1 will generate the corresponding centre density map with size of 128×128.

each of the microatolls has more significant and features in drone images. Therefore, G_1 is able to distinguish and indicate those areas where there could be microatolls via adjusting the receptive field to 50×50 without the use of dilated convolutional layer. With the same convolutional building blocks described in D, the structure of G_1 is shown in Fig. 4.4. An input image serially goes through ten convolutional building blocks, and G_1 outputs a density map $\mathbf{S}$ that is four times smaller than $\mathbf{I}$ and each of the eight connected components in $\mathbf{S}$ is a detected microatoll. An eight connected component in our density map is the maximum set of pixels in the image such that any two pixels in the set are eight connected.

In terms of G_2, it aims to reduce the FA by cooperating with the context information. As the experimental study discussed in Chapter 3, the context information will improve the microatoll detection performance. Coincidentally, a lager receptive field in G_2 encloses context information performing a similar manner. However, the original receptive field is not suitable for the microatoll detection task because most of the bounding boxes are below 75 in both width and height, and they are far smaller than the receptive field in G_2 (258×258). Therefore, the receptive field of G_2 is reduced to 175×175. Selecting

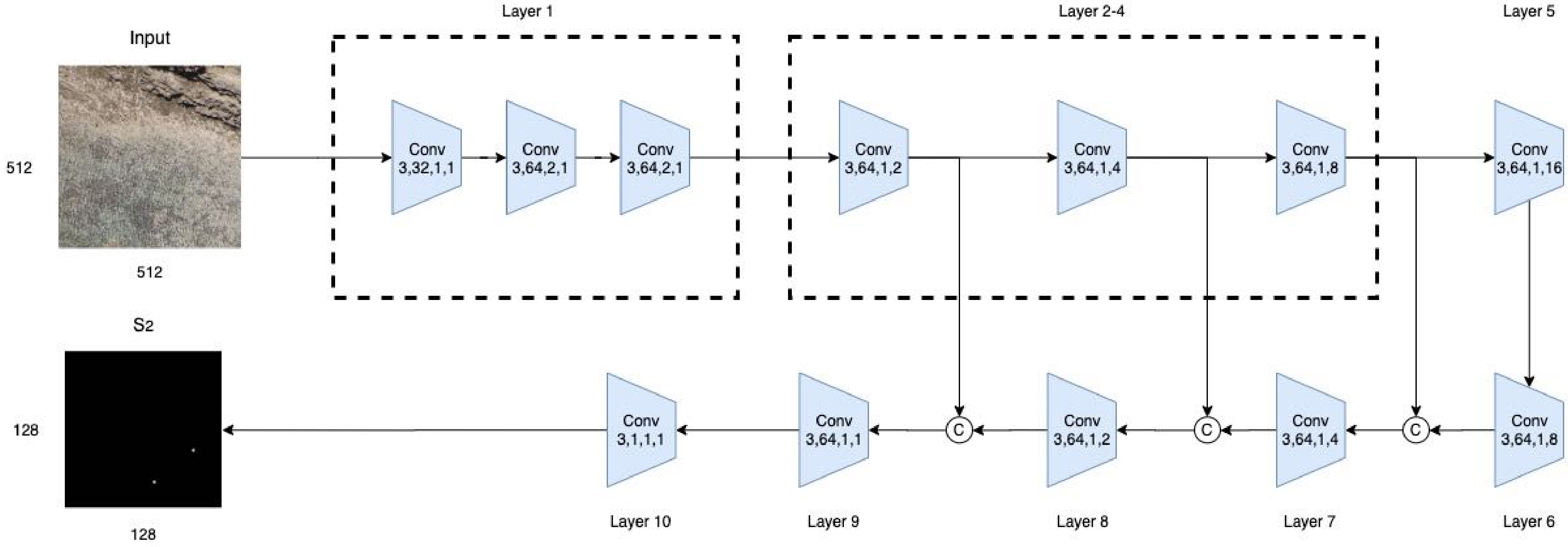


Figure 4.5: The structure of the generator which is used to reduce false alarm, named G_2.

such a receptive field size is because i) our targets are individual microatolls that are well separated, ii) most targets are under the size of 75 × 75, so the receptive field is enlarged two times bigger than the size to enclose rich context information, and iii) there are few microatolls in size of 175 × 175 pixels, so the receptive field is enlarged to meet this size. According to ResNet [34], residual networks are easier to be optimized and can gain accuracy from considerably increased depth. Hence, three residual shortcuts are built in G_2. The reason behind only using residual structure in G_2 is simple. G_1 aims to distinguish those small patches with the highest possibility to be microatolls, and most of the microatolls are salient objects within the 75 × 75 bounding boxes, so a shallow network is sufficient to extract the local features. While G_2 cooperates with context information to remove those patches that are not microatolls, the additional context information needs a deeper network to handle the feature extraction process. As shown in Fig. 4.5, an input image **I** goes through twelve convolutional building blocks to generate a density map that is similar to the outputs of G_1. Meanwhile, there are residual shortcuts in the pairs of 2^{nd} and 8^{th}, 3^{rd} and 7^{th} ,and 4^{th} and 6^{th}.

4.4 Experimental Result

4.4.1 Evaluation Method

Gao et al. [68] use mean absolute error $MAE = \frac{1}{N}\sum_{i=1}^{n}|y-\hat{y}|$ and root mean square error $RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(y-\hat{y})^2}$ to evaluate their model performance in counting object tasks. The outputs in the counting object tasks are similar to ours, so MAE and RMSE are used to evaluate MDFA and Microatoll-GAN originally. However, it is soon realized that these two methods do not meet our requirement. Similar to the MD and FA in literature, MAE and RMSE strictly compute the distance between binary y and $\hat{y}$. The distance will increase when the detected centre does not exactly match the ground truth. However, in our detection setting, they are still considered correct if they are closely around the microatoll centres.

Accordingly, the true positive and false positive concepts are employed to propose an evaluation method as follows. First, the eight connected components in detected density maps will be counted as detected microatolls. Then, each of the connected components will be compared with original bounding boxes. Furthermore, the connected-component is a true positive as long as it is within a bounding box; otherwise, it is a false positive. Second, it is found that most individual microatolls are well separated during the labelling process, which means the overlapped bounding boxes rarely appear in the microatoll dataset. Therefore, if there are more than one connected components in a bounding box, only the biggest one is treated as the true positive, and the rest are treated as the false positive results. In sum, the evaluation method is described as follows:

- All eight connected components are regarded as detected targets.
- True positive detection results are those connected-components within original bounding boxes.
- False positive detection results are those outside the original bounding boxes or those connected-components smaller than the biggest one in the same bounding box.

4.4.2 Experimental Setting

All experiments are implemented in PyTorch [69], and our model is trained from scratch in a server with two NVIDIA 1080ti GPUs. As described in Section 4.3.2, all input images are cropped into 512×512 patches. In the training set, the stride is set as 100 to enlarge the training samples. Different to the training set, the stride in the testing set (only taking G123 for testing, see Section 3.3) is 512. Consequently, a microatoll will only appear once in the test dataset. Based on this setting, performance evaluation becomes more reliable.

To train the model, Adam [70] is used to optimize D, G_1 and G_2 by setting the learning rates as $1e^{-4}$, $1e^{-5}$ and $1e^{-5}$, respectively. The model performs in an end-to-end manner. The setting of hyper-parameters in the final loss function (see Equation 4.2) are $\alpha_1 = 1, \alpha_2 = 1, \alpha_3 = 100$, $\lambda_1 = 10$ and $\lambda_2 = 5$ for 10 epochs without other training strategies.

4.4.3 Result Analysis

Following the evaluation method described previously, the comparison of the model performances on microatoll detection task is shown in Table 4.1. In total, there are 515 unique microatolls in the test dataset. The original MDFA could only detect 129 microatolls. In contrast, our proposed Microatoll-GAN increases the true positive to 284. More details are given in Table 4.1 and Fig. 4.6. There is a significant improvement in both recall and precision.

Model	Ground Truth	Detected	True Positive	False Positive	Recall	Precision
MDFA [13]	515	231	129	98	0.25	0.57
M-GAN (ours)		447	284	162	0.57	0.64

Table 4.1: The performances of the MDFA and M-GAN in G123. There is a significant increase from 0.25 to 0.57 in the recall, and the precision is elevated too.

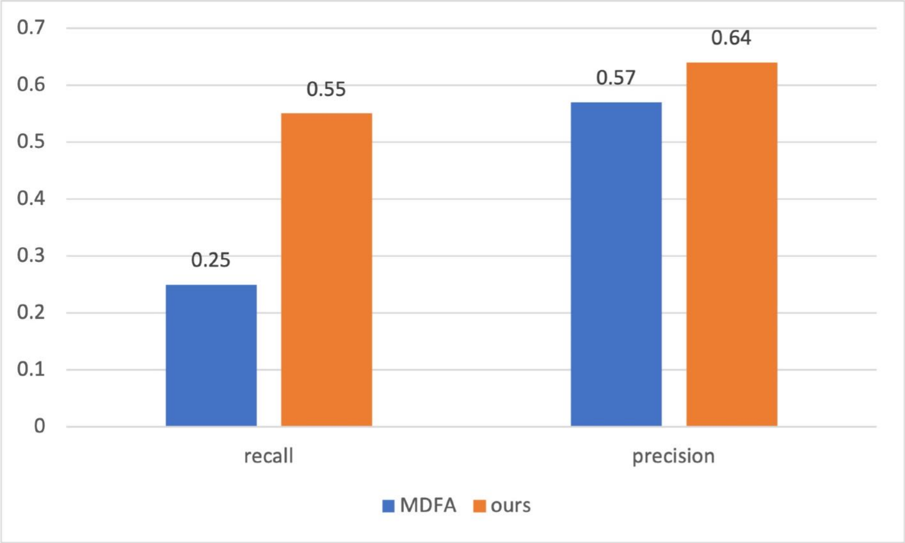


Figure 4.6: The Recall and Precision of MDFA and M-GAN on the test group. The significant improvement is obtained compared with original MDFA [13]. Best viewed in color.

The increase in numbers cannot fully reflect the actual performance. To figure out the reasons for performance enhancement, those typical areas that give better results are selected and, at the same time, those areas that have overall performance dropped will be presented as well. Fig. 4.7 illustrates the typical areas where our model can perform well.

Column (a) gives the samples of input images, and green box labels indicate all ground truth. From top to bottom, the first image indicates the area containing a dense microatoll but with significant and rich features. Each of the microatolls has a significant margin and well separated from others. The second one shows the microatolls with various sizes, and the last one indicates the microatolls appearing in the surrounding where background has similar color with microatolls. Column (b) to (d) illustrate the ground truth, the results from MDFA and Microatoll-GAN (ours), respectively. In the first row, the proposed Microatoll-GAN detects all microatolls and MDFA roughly detects half of them. Similarly, our model surpasses MDFA in other areas. The comparison highlights that our proposed method can better handle the microatoll centre detection task. Simultaneously, the previous MDFA cannot locate the microatolls even in the area with significant features and the same size (see the first row). Based on this, the proposed model indeed improves the performance compared to the original MDFA.

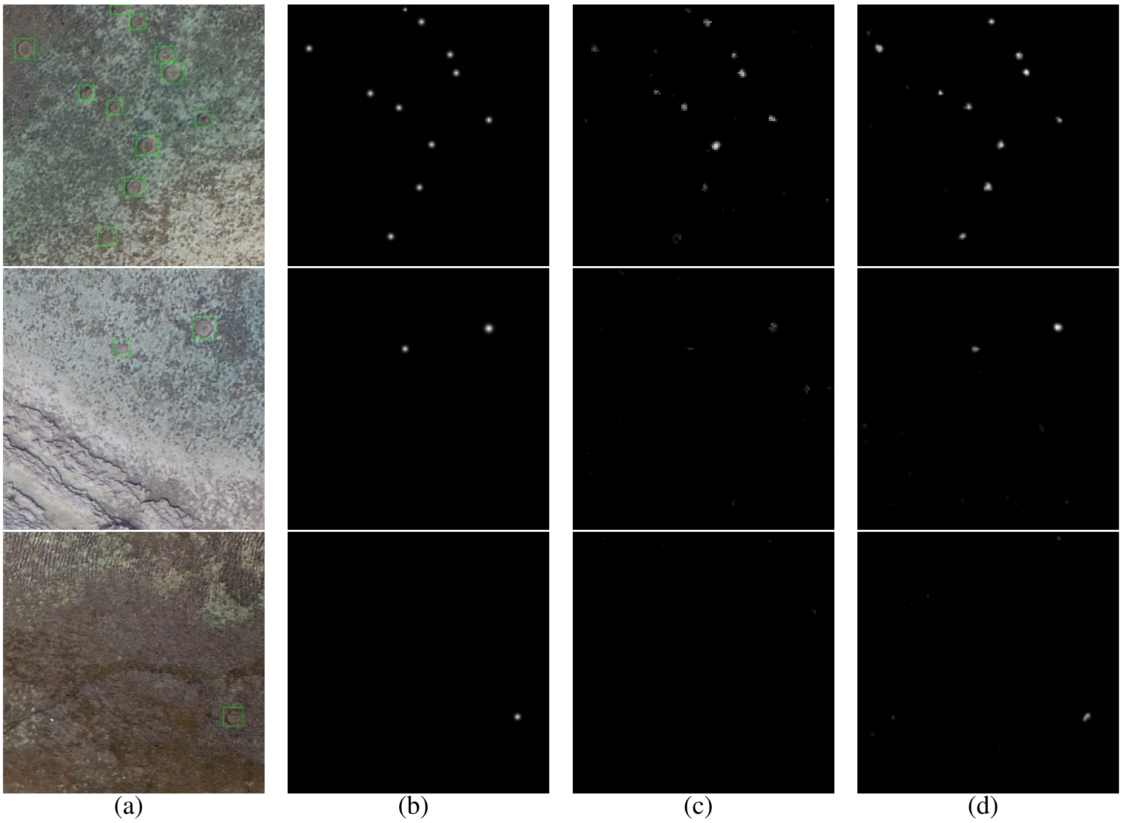


Figure 4.7: There are three examples to indicate those areas which are well detected by the proposed method. Column (a) is the input images, and the green boxes are to show the microatolls in the area. Column (b) is the ground truths heatmap, (c) is generated by the MDFA [13] and (d) is the result coming from our model.

Three main factors affecting the detection performance are: water wave, microatoll breakage, and the target sizes unsuitable for the fixed receptive field. Those areas cannot be well detected as presented in Fig. 4.8. The first row indicates that the water wave covered microatoll is not detected in both models. Second, as described in Section 4.3.2, the receptive fields in G_1 and G_2 are fixed. The fixed receptive fields perform well in most cases, but those microatolls that are extremely smaller or bigger than the receptive fields will be considered background by mistake. Third, the breakage will be another important factor in reducing performance. As shown in the third row, our model cannot correctly detect the microatoll even it has a significant margin. Similarly, the one on the right side is wrongly taken as background for the same reason.

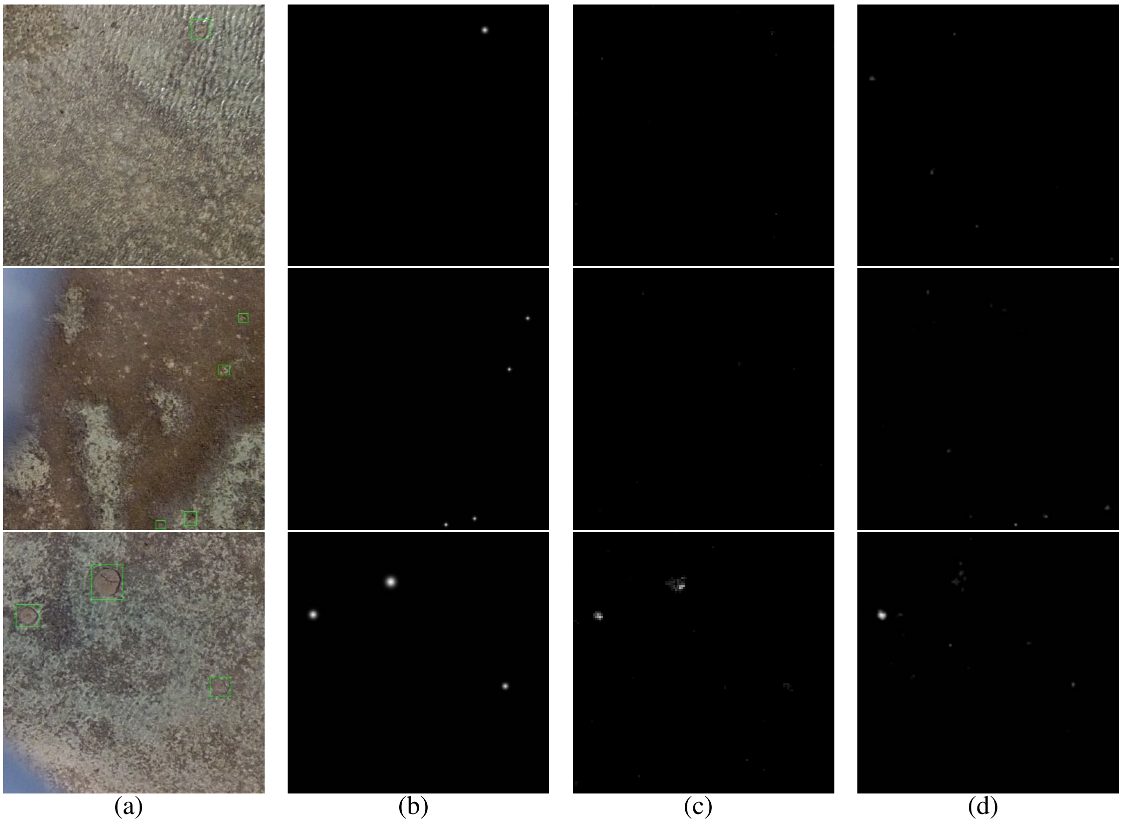


Figure 4.8: There are three examples to indicate the reasons for the dropped detection performances. Water wave impact is shown in the first row, the microatoll size unsuitable issues is in the second row and the last row shows the impact of microatoll breakage.

4.5 Chapter Summary

This chapter takes advantages of MDFA that separates the reduction of missed detection and false alarm into two models. As mentioned before, although two generators have different purposes, where the G_1 is to reduce the MD and the G_2 is to reduce the FA, the information exchange is performed by L_{MF} to boost the ability of G_1 (designed to reduce MD) in reducing FA and G_2 (designed to reduce FA) in reducing MD. The information exchange is as shown in Equation 4.4, the item MD and FA appear in L_{MF_1} and L_{MF_2} with different λ. That means the two sub-task share the information but have different focus. At the same time, L_{cGAN} helps to force G_1 and G_2 to produce the same feature maps and align the outputs to the ground truth.

The loss function, models, and dataset are modified to bridge the gap between the infrared small object segmentation and the microatoll detection from drone images. The hyper-parameters in the loss function are adjusted by experimental study. In terms of model, original model outputs have the same size as input. This setting may benefit from producing pixel-level targets, but it will bring a high computation cost when the input images have a larger size. The cost is reduced by setting the stride as two in the convolutional layer in generators and employing the average pooling layer in the discriminator to downsample the input images. Because the training annotations are transformed from bounding boxes in our microatoll dataset, it is too strict to using only one or two pixels to represent the location of microatoll centres. Therefore, a Gaussian kernel is used to reconstruct the ground truth of our microatoll dataset for the proposed Microatoll-GAN. After that, an experimental study is conducted to compare MDFA and Microatoll-GAN performances, and our method surpasses MDFA in both recall and precision.

The advantages of the proposed Microatoll-GAN are significant. It removes the need for anchors and the post-processing step. Meanwhile, our model is trained end-to-end from scratch. It successfully converts a segmentation model to meet the requirements of microatoll detection task. In terms of disadvantages, our model is still heavily affected by water wave, different microatoll sizes and the breakages. To solve these problems, further research could focus on combining FPN with Microatoll-GAN and enhancing the generalisation of the model.

Chapter 5

Conclusion

Drones are increasingly used in field research for microatoll image collection. Hence, automatically locating their positions becomes a key step in making more effective use of the captured drone images. To the best of our knowledge, this thesis is the first one that systematically applies deep learning in detecting microatolls from drone images. Therefore, dataset construction is the first critical challenge. As the first contribution, microatoll characteristics are investigated to establish microatoll labelling rules for microatoll annotation creation and construct the first drone image based microatoll detection dataset. Meanwhile, a set of experimental studies is presented to compare the performances of two-stage detection methods (i.e. Faster R-CNN) and one-stage detection methods (i.e. SSD), evaluates enhancements from the detection modules such as FPN and OHEM, and highlights the challenges that appear in the microatoll detection task. To solve one of the challenges that the microatoll pavements drop the detection performance in the most challenging area (G4), Microatoll-Net is proposed in Chapter 3 to utilise additional context information outside the RoIs to distinguish individual microatolls and microatoll pavements. Besides, variants Microatoll-Net are developed and evaluated to address the best way to combine local and context features.

The use of anchors and post-processing steps are widespread in modern detection methods. In Chapter 4, Microatoll-GAN is proposed to remove the need for anchors and post-processing steps. There are two generators in Microatoll-GAN to reduce miss detection and false alarm, and a discriminator helps to align two density maps with the ground truth and minimise the difference of feature maps between G_1 and G_2. The proposed microatoll dataset is re-constructed, which presents the regression targets as two-dimension Gaussian density maps instead of bounding boxes. The proposed Microatoll-GAN significantly outperforms the MDFA in our microatoll dataset.

5.1 Future Research Directions

This thesis emphasises the challenges of detecting microatolls from drone images: areas not seen in our microatoll dataset(i.e. mangrove forest and helipad), tides and sunshine impact microatolls characters, and microatoll pavements. The proposed Microatoll-Net in Chapter 3 partially addresses the issue of microatoll pavements, but it can still be improved. Meanwhile, more than 2,400 islands are in the Great Barrier Reef region. Maintaining or improving the performance of the proposed Microatoll-Net to detect microatolls from the drone images of other islands is also a significant issue for future research. As for the adversarial learning network based Microatoll-GAN proposed in Chapter 4, its performance is considerably affected due to the varying dimensions of microatolls which are far beyond its fixed receptive field of the proposed model. In other words, the microatolls that are too small or too big for the receptive fields of the proposed model will be wrongly recognised as background. Also, the water wave hurts detection performance. Therefore, improving the generalisation capability of the proposed model and handling the problem of various sizes presented in microatolls are the potential directions to enhance detection performance. Additionally, applying the Microatoll-GAN to common object detection tasks will also be an exciting research direction to explore because of the advantages of removing anchors and post-processes. Removing anchors will be an important step to make computer vision to be close to human vision behaviour because human being will not set an anchor for locating objects. Meanwhile, it is also simplify the algorithm design process. Besides, removing post-processes will make the model more robust because the results are no longer dependent on the experience of experts.

Bibliography

(1) C. D. Woodroffe, H. V. McGregor, K. Lambeck, S. G. Smithers and D. Fink, “Mid-Pacific microatolls record sea-level stability over the past 5000 yr”, *Geology*, 2012, **40**, 951–954.

(2) T. P. Scoffin, D. R. Stoddart and B. R. Rosen, “The Nature and Significance of Microatolls”, *Philosophical Transactions of the Royal Society B: Biological Sciences*, 1978, **284**, 99–122.

(3) A. Krizhevsky, I. Sutskever and G. E. Hinton, “ImageNet classification with deep convolutional neural networks”, *Advances in Neural Information Processing Systems*, 2012.

(4) Y. LeCun, L. Bottou, Y. Bengio and P. Haffner, “Gradient-based learning applied to document recognition”, *Proceedings of the IEEE*, 1998, **86**, 2278–2323.

(5) N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection”, *Proceedings - 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005*, 2005, **I**, 886–893.

(6) T. Lindeberg, “Scale Invariant Feature Transform”, *Scholarpedia*, 2012, DOI: `10.4249/scholarpedia.10491`.

(7) D. Du, P. Zhu, L. Wen, X. Bian, H. Lin, Q. Hu, T. Peng, J. Zheng, X. Wang, Y. Zhang, L. Bo, H. Shi, R. Zhu, A. Kumar, A. Li, A. Zinollayev, A. Askergaliyev, A. Schumann, B. Mao, B. Lee, C. Liu, C. Chen, C. Pan, C. Huo, D. Yu, D. Cong, D. Zeng, D. R. Pailla, D. Li, D. Wang, D. Cho, D. Zhang, F. Bai, G. Jose, G. Gao, G. Liu, H. Xiong, H. Qi, H. Wang, H. Qiu, H. Li, H. Lu, I. Kim, J. Kim, J. Shen, J. Lee, J. Ge, J. Xu, J. Zhou, J. Meier, J. W. Choi, J. Hu, J. Zhang, J. Huang, K.

Huang, K. Wang, L. Sommer, L. Jin, L. Zhang, L. Huang, L. Sun, L. Steinmann, M. Jia, N. Xu, P. Zhang, Q. Chen, Q. Lv, Q. Liu, Q. Cheng, S. S. Chennamsetty, S. Chen, S. Wei, S. S. S. Kruthiventi, S. Hong, S. Kang, T. Wu, T. Feng, V. A. Kollerathu, W. Li, W. Dai, W. Qin, W. Wang, X. Wang, X. Chen, X. Chen, X. Sun, X. Zhang, X. Zhao, X. Zhang, X. Zhang, X. Chen, X. Wei, X. Zhang, Y. Li, Y. Chen, Y. H. Toh, Y. Zhang, Y. Zhu, Y. Zhong, Z. Wang, Z. Wang, Z. Song and Z. Liu, “VisDrone-DET2019: The Vision Meets Drone Object Detection in Image Challenge Results”, 2019, 213–226.

(8) M. Everingham, S. M. Ali Eslami, L. Van Gool, C. K. I Williams, J. Winn, A. Zisserman, M. Everingham, S. M. A Eslami, J. Winn, L. K. Van Gool Leuven, B. L. Van Gool ETH, S. C. K I Williams and A. Zisserman, “The PASCAL Visual Object Classes Challenge: A Retrospective”, *Int J Comput Vis*, 2015, **111**, 98–136.

(9) T.-Y. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, P. Dollar and L. Zitnick, “Microsoft COCO: Common Objects in Context”, 2014, **8693 LNCS**, 740–755.

(10) F. Johnson, “A Bibliography of Radiocarbon Dating”, *Radiocarbon*, 1959, **1**, 199–214.

(11) S. Ren, K. He, R. Girshick and J. Sun, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, **39**, ed. C. Cortes, N. Lawrence, D. Lee, M. Sugiyama and R. Garnett, 1137–1149.

(12) W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu and A. C. Berg, “SSD: Single Shot MultiBox Detector”, *European conference on computer vision*, 2016, **32**, 21–37.

(13) H. Wang, L. Zhou and L. Wang, “Miss Detection vs. False Alarm: Adversarial Learning for Small Object Segmentation in Infrared Images”, *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, 8508–8517.

(14) P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features”, *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2001, DOI: `10.1109/cvpr.2001.990517`.

(15) Y. Freund and R. E. Schapire, “A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting”, *Journal of Computer and System Sciences*, 1997, **55**, 119–139.

(16) Y. LeCun, C. Cortes and C. Burges, “MNIST handwritten digit database”, *ATT Labs [Online].*

(17) P. Felzenszwalb, D. McAllester and D. Ramanan, “A discriminatively trained, multiscale, deformable part model”, *26th IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2008, DOI: `10.1109/CVPR.2008.4587597`.

(18) R. Girshick, J. Donahue, T. Darrell and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation”, *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2014, 580–587.

(19) K. He, X. Zhang, S. Ren and J. Sun, “Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015, **37**, 1904–1916.

(20) R. Girshick, “Fast R-CNN”, *Proceedings of the IEEE International Conference on Computer Vision*, 2015, **2015 Inter**, 1440–1448.

(21) T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan and S. Belongie, “Feature Pyramid Networks for Object Detection”, *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, **2017-Janua**, 936–944.

(22) J. Redmon, S. Divvala, R. Girshick and A. Farhadi, “You Only Look Once: Unified, Real-Time Object Detection”, *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, 779–788.

(23) J. Redmon and A. Farhadi, “YOLO9000: Better, Faster, Stronger”, *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, **2017-Janua**, 6517–6525.

(24) J. Redmon and A. Farhadi, “YOLOv3: An Incremental Improvement”, *arXiv preprint arXiv:1804.02767*, 2018.

(25) A. Bochkovskiy, C.-Y. Wang and H.-Y. M. Liao, *YOLOv4: Optimal Speed and Accuracy of Object Detection*, tech. rep., 2020.

(26) T.-Y. Lin, P. Goyal, R. Girshick, K. He and P. Dollar, “Focal Loss for Dense Object Detection”, *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, **20**, 2999–3007.

(27) H. Law and J. Deng, “CornerNet: Detecting Objects as Paired Keypoints”, *Anticancer research*, 2018, **24**, 765–781.

(28) K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang and Q. Tian, “CenterNet: Keypoint Triplets for Object Detection”, *arXiv preprint arXiv:1904.08189*, 2019.

(29) X. Zhou, D. Wang and P. Krähenbühl, “Objects as Points”, *arXiv preprint arXiv:1904.07850*, 2019.

(30) T. Kong, F. Sun, H. Liu, Y. Jiang and J. Shi, “FoveaBox: Beyond Anchor-based Object Detector”, *arXiv preprint arXiv:1904.03797*, 2019.

(31) Z. Tian, C. Shen, H. Chen and T. He, “FCOS: Fully Convolutional One-Stage Object Detection”, *arXiv preprint arXiv:1904.01355*, 2019.

(32) N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov and S. Zagoruyko, “End-to-End Object Detection with Transformers”, *arXiv preprint arXiv:2005.12872*, 2020.

(33) V. Nair and G. E. Hinton, “Rectified Linear Units Improve Restricted Boltzmann Machines”, *Proceedings of the 27th International Conference on International Conference on Machine Learning*, 2010, 807–814.

(34) K. He, X. Zhang, S. Ren and J. Sun, “Deep Residual Learning for Image Recognition”, *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, **2016-Decem**, 770–778.

(35) K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition”, *International Conference on Learning Representations*, 2015.

(36) J. Redmon, *Darknet: Open Source Neural Networks in C.*

(37) C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke and A. Rabinovich, "Going deeper with convolutions", *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2015, **07-12-June**, 1–9.

(38) J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers and A. W. M. Smeulders, "Selective Search for Object Recognition", *International Journal of Computer Vision*, 2013, **104**, 154–171.

(39) K. He, G. Gkioxari, P. Dollar and R. Girshick, "Mask R-CNN", *Proceedings of the IEEE International Conference on Computer Vision*, 2017, **2017-Octob**, 2980–2988.

(40) J. Pang, K. Chen, J. Shi, H. Feng, W. Ouyang and D. Lin, "Libra R-CNN: Towards Balanced Learning for Object Detection", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

(41) Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into High Quality Object Detection", *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018, 6154–6162.

(42) Z. Yao, Y. Cao, S. Zheng, G. Huang and S. Lin, "Cross-Iteration Batch Normalization", *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, 8759–8768.

(43) S. Liu, L. Qi, H. Qin, J. Shi and J. Jia, "Path Aggregation Network for Instance Segmentation", *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018, DOI: `10.1109/CVPR.2018.00913`.

(44) S. Woo, J. Park, J.-Y. Lee and I. S. Kweon, "CBAM: Convolutional Block Attention Module", *ECCV*, 2018, **11211 LNCS**, 3–19.

(45) A. Newell, K. Yang and J. Deng, "Stacked hourglass networks for human pose estimation", *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2016, **9912 LNCS**, 483–499.

(46) H. Law, Y. Teng, O. Russakovsky and J. Deng, “CornerNet-Lite: Efficient Keypoint Based Object Detection”, *arXiv preprint arXiv:1904.08900*, 2019.

(47) A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, “Attention Is All You Need”, *Advances in Neural Information Processing Systems*, 2017.

(48) P. Mittal, R. Singh and A. Sharma, “Deep learning-based object detection in low-altitude UAV datasets: A survey”, *Image and Vision Computing*, 2020, **104**, 104046.

(49) B. Kellenberger, D. Marcos and D. Tuia, “Detecting Mammals in UAV Images: Best Practices to address a substantially Imbalanced Dataset with Deep Learning”, 2018, DOI: `10.1016/j.rse.2018.06.028`.

(50) J. Pang, C. Li, J. Shi, Z. Xu and H. Feng, “R^2-CNN: Fast Tiny Object Detection in Large-Scale Remote Sensing Images”, *arXiv preprint arXiv:1902.06042*, 2019, DOI: `10.1109/TGRS.2019.2899955`.

(51) S. Hamylton, R. Morris, R. Carvalho, N. Roder, P. Barlow, K. Mills and L. Wang, “Evaluating techniques for mapping island vegetation from unmanned aerial vehicle (UAV) images: Pixel classification, visual interpretation and machine learning approaches”, *International Journal of Applied Earth Observation and Geoinformation*, 2020, **89**, 102085.

(52) P. Zhu, L. Wen, X. Bian, H. Ling and Q. Hu, “Vision Meets Drones: A Challenge”, *arXiv preprint arXiv:1804.07437*, 2018.

(53) A. Shrivastava, A. Gupta and R. Girshick, “Training region-based object detectors with online hard example mining”, *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016, **2016-Decem**, 761–769.

(54) S. Bell, C. L. Zitnick, K. Bala and R. Girshick, “Inside-Outside Net: Detecting Objects in Context with Skip Pooling and Recurrent Neural Networks”, *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, 2874–2883.

(55) Q. V. Le, N. Jaitly and G. E. Hinton, “A Simple Way to Initialize Recurrent Networks of Rectified Linear Units”, *arXiv preprint arXiv:1504.00941*, 2015.

(56) X. Tang, D. K. Du, Z. He and J. Liu, “PyramidBox: A context-assisted single shot face detector”, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2018, **11213 LNCS**, 812–828.

(57) K. Kuan, G. Manek, J. Lin, Y. Fang and V. Chandrasekhar, “Region average pooling for context-aware object detection”, *Proceedings - International Conference on Image Processing, ICIP*, 2018, **2017-Septe**, 1347–1351.

(58) L. Wang, J. Zhang, L. Zhou, C. Tang and W. Li, “Beyond covariance: Feature representation with nonlinear kernel matrices”, *Proceedings of the IEEE International Conference on Computer Vision*, 2015, **2015 Inter**, 4570–4578.

(59) G. Zilin, X. Jiangtao, W. Qilong and L. Peihua, “Global Second-order Pooling Convolutional Networks”, *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

(60) T. Lin, *labelImg*, `https://github.com/tzutalin/labelImg`.

(61) In, *Encyclopedia of Machine Learning*, ed. C. Sammut and G. I. Webb, Springer US, Boston, MA, 2010, pp. 600–601.

(62) O. J. Oyelade, O. O. Oladipupo and I. C. Obagbuwa, *Application of k-Means Clustering algorithm for prediction of Students' Academic Performance*, tech. rep. 1, 2010.

(63) P. Isola, J. Y. Zhu, T. Zhou and A. A. Efros, “Image-to-image translation with conditional adversarial networks”, *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017, DOI: `10.1109/CVPR.2017.632`.

(64) X. Zhou, J. Zhuo and P. Krahenbuhl, “Bottom-Up Object Detection by Grouping Extreme and Center Points”, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

(65) F. Yu and V. Koltun, “Multi-Scale Context Aggregation by Dilated Convolutions”, *arXiv preprint arXiv:1511.07122*, 2015.

(66) M. Arjovsky and L. Bottou, “Towards Principled Methods for Training Generative Adversarial Networks”, *arXiv preprint arXiv:1701.04862*, 2017.

(67) B. Xu, N. Wang, T. Chen and M. Li, “Empirical Evaluation of Rectified Activations in Convolutional Network”, *arXiv preprint arXiv:1505.00853*, 2015.

(68) G. Gao, Q. Liu and Y. Wang, *IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING 1 Counting from Sky: A Large-scale Dataset for Remote Sensing Object Counting and A Benchmark Method*, tech. rep.

(69) A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai and S. Chintala, “PyTorch: An imperative style, high-performance deep learning library”, 2019.

(70) D. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization”, *Computer Science*, 2014.