CHAPTER I
INTRODUCTION

1.1 Background

Remote sensing images are utilized in a variety of fields, such as traffic monitoring [1], urban planning [2], intelligent monitoring [3], and environmental monitoring [4]. Advancements in deep learning technology have analysed remote sensing images more efficiently and intelligently, minimizing the reliance on manual processes. As essential tasks in computer vision, object recognition and detection are crucial for extracting meaningful insights from these images [1].

In remote sensing, images are captured using various technologies such as satellites, aircraft, and, more recently, drone-based sensors. As a result, gathering training data for remote sensing is more challenging. However, the unique remote sensing image capture methods introduce several vital challenges. Firstly, objects within these images are often relatively small compared to the overall image size, which limits the amount of feature information available for detection. For instance, an object might only cover a few dozen pixels in an image with millions of pixels. Second, because objects might appear at different angles, object detection algorithms must provide characteristics invariant to rotation to improve detection accuracy [5]. The size of objects can vary significantly both within a one image and divergent images, such as the difference in scale between a car and an airplane [6], [7], [8], [9], [10], [11]. Object detection algorithms must also address noise and occlusions [12].

Deep learning techniques have proven to be highly effective in this domain and have grown to be an important subject of study [13], [14]. These techniques generally fall into two categories: supervised and unsupervised. Unsupervised learning involves extracting patterns or high-level semantics from unlabelled data. In contrast, Convolutional Neural Networks (CNNs) have become a preferred choice for supervised object detection across various image processing domains. Popular You Only Look Once (YOLO) models have demonstrated outstanding performance [15], [16], [17], as they demonstrate a strong capability to simultaneously learn regions of interest and contextual information. You Only Look Once version 8 (YOLOv8) [18] has demonstrated efficiency and accuracy, achieving high mean average precision. Even though it does not explicitly target small object detection, remains a robust benchmark method [19], [20].

Remote sensing images frequently exhibit significant variances in target scale and are influenced by complicated environmental elements such as lighting and weather. These challenges make it difficult to achieve optimal performance when applying standard deep learning models directly to these images [21]. In response to these challenges, researchers often adjust R-CNN and YOLO-based models to meet the specific requirements of remote sensing applications [22]. Xu et al. [23] developed an improved remote sensing object detection framework on the basic of You Only Look Once version 3 (YOLOv3). This framework incorporates a feature enhancement network designed to extract essential features more effectively. When tested on remote sensing datasets, the model successfully detected targets.

Liu et al. [24] proposed an improved model on the basic of YOLOv8, called YOLO-SSP, which enhances detection accuracy by optimizing the downsampling layers to capture finer details and utilizing hierarchical pooling operations to generate weights from different spatial locations. Despite various efforts to enhance multi-scale object detection, accurately identifying tightly clustered small targets in remote sensing imagery remains a significant challenge.

Liu et al. [25] introduced the YOLO-extract method, inspired by You Only Look Once version 5 (YOLOv5), which utilizes residual concepts to enhance feature extraction capabilities. This model integrates the coordinate attention mechanism and mixed dilated convolution. To speed up model convergence, Focal-α EIoU was used to replace the CIoU loss function. Additionally, an extra detection head was incorporated to specifically target small and densely packed objects.

Zhang et al. [26] enhanced the backbone network's receptive field and feature fusion components based on YOLOv5, improving detection performance by strengthening the feature representation of small objects. However, enhanced multiscale feature extraction typically results in more complex model architectures and increased computational demands, which can reduce the model's inference speed. Some approaches aim to increase the model's consideration to small objects by refining the distance metrics or adjusting the bounding box thresholds.

Recent advancements in deep learning have led to the development of highly efficient and accurate models, with YOLOv8 [18] emerging as one of the state-of-the-art (SOTA) methods. YOLOv8 has demonstrated impressive performance, achieving high mean average precision on benchmark datasets [19], [20]. However, while it excels in detecting medium to large objects, the model encounters challenges when applied to small objects, which require specialized techniques for optimal detection.

YOLOv8's architecture, including using the C2f backbone, is effective for handling larger objects, but it introduces redundancy in feature extraction, reducing computational efficiency. Additionally, the network's neck, responsible for connecting the backbone to the detection head, suffers from issues such as vanishing gradients and inadequate feature fusion, which hinder the model's ability to capture fine-grained details crucial for detecting small objects. Furthermore, the detection head lacks specific adaptations that enhance its ability to detect small-scale objects, leading to reduced accuracy when applied to such tasks.

Moreover, YOLOv8 employs the Complete Intersection over Union (CIoU) loss function, which is well-suited for detecting standard-sized objects. However, CIoU struggles with small objects, where the discrepancies between predicted and ground truth bounding boxes are minimal. In such cases, CIoU's performance couldn’t be better, as it is less effective at handling the subtle differences characteristic of small object detection.

This research addresses these limitations by proposing modifications to YOLOv8's architecture, explicitly targeting the backbone, neck, and detection head to improve its capability in detecting small objects. In addition, explore alternative loss functions better suited for small object detection, aiming to enhance model performance in scenarios where objects of interest occupy fewer pixels or are partially occluded. The findings of this study aim to contribute to the advancement of object detection models by improving their efficiency and accuracy in detecting small objects.

This research presents a proposed YOLOv8-based object detection architecture for remote sensing images trained on the DIOR dataset. This research outlines the key contributions as follows:

1. Firstly, a special C2f layers in the backbone network has been replaced by the DCN_C2f module [4], [27], which uses deformation convolution to better capture the features of objects with complex shapes and patterns. This enables the model to be more adaptive in recognizing objects with irregular contours and various visual variations. Subsequently, a more sophisticated attention mechanism was integrated into the neck part of the YOLOv8n model by incorporating ResBlock_CBAM [28], [29] during the training process. This mechanism enables the model to focus more on core features, thus improving the precision of object detection.
2. Secondly, a high-resolution head was added to the head of the model, which significantly improved the ability to detect smaller objects. This change makes the model more sensitive to subtle details that are often overlooked by conventional detection methods.
3. Finally, the loss function was optimised using the Distance Intersection over Union (DIoU) [30] approach to improve the overall performance of the model. This approach emphasizes the modelling of spatial relationships between predicted and ground truth boxes, improving accuracy and efficiency across various detection settings.

1.2 Problem Identification

Object detection in remote sensing presents challenges, particularly in detecting small, occluded, or multi-scale objects. This often results in missed detections or false positives in conventional models like YOLOv8 feature extraction and fusion limitations, especially for small objects. The fixed receptive field of standard convolutions and inefficient feature fusion processes contribute to poor detection accuracy, particularly for small-scale targets. Additionally, the current configuration of three detection heads in YOLOv8 must be improved to handle a wide range of object sizes, including small targets.

This research explores advanced loss functions Generalized Intersection over Union (GIoU), DIoU, CIoU, Scylla Intersection over Union (SIoU), and Wise Intersection over Union (WIoU) to enhance bounding box regression and improve the detection of small objects. These loss functions provide better spatial alignment, scale consistency, and adaptive weighting for various object sizes, addressing the limitations of traditional IoU-based methods. The aim is to enhance model accuracy, particularly for small, occluded, or multi-scale objects, by optimizing the regression process and improving overall detection performance.

1.3 Objectives and Contributions

This thesis is based on the following key assumptions:

1. The study proposes modifications to the YOLOv8 architecture, explicitly targeting the backbone, neck, and head components.
2. It seeks to optimize the loss function for improved performance.
3. The performance of the proposed model will be compared against existing models to evaluate its effectiveness.

1.4 Scope of Work

To ensure this thesis aligns with the stated requirements, it does not appear to modify the subject or scope of work.

1. The system will be implemented using Python for its development.
2. The base algorithm is YOLOv8, with modifications applied to the backbone and head networks.
3. The objects to be detected comprise 20 classes, all located on the ground.
4. The dataset is DIOR, proposed by KeLi et al. [3].
5. Data processing includes converting annotations from .xml to YOLOv8 format (.txt).
6. Precision, Recall, F1-Score, and mAP@0.5IoU metrics will assess the model's performance.

1.5 Expected Results

Before experimenting, it is expected that the proposed EXYOLOv8-Exploration2 model, combined with the DIoU loss function and CSPDarknet backbone, will significantly enhance object detection performance regarding accuracy, efficiency, and robustness. The model is anticipated to achieve higher mean Average Precision (mAP) scores compared to existing methods like GIoU, CIoU, SIoU, and WIoU, primarily because of DIoU's ability to directly minimize the distance between the center points of predicted and ground truth bounding boxes. This focus is expected to result in better localization and spatial alignment of detected objects.

1.6 Research Methodology

This thesis employs fundamental research and experimental methods structured around distinct work packages (WP). The work packages for this thesis are outlined as follows:

- **WP 1**: Conducting a comprehensive literature review.
- **WP 2**: Selecting and configuring the model.
- **WP 3**: Converting the dataset format from .xml to .txt for YOLOv8 format.
- **WP 4**: Modifying the backbone by replacing the C2f layer with DCN_C2f.
- **WP 5**: Enhancing the neck by adding a ResBlock_CBAM layer.
- **WP 6**: Implementing modifications to the head of the network.
- **WP 7**: Validating the model using test data.
- **WP 8**: Evaluating the model's performance and assessing accuracy improvements using Precision, Recall, F1-Score, and mAP@0.5IoU metrics.

1.7 Structure of Thesis

The following is the structure of this thesis:

- **CHAPTER II: BASIC CONCEPT**

  This chapter provides an overview of this thesis's fundamental concepts and algorithms.
- **CHAPTER III: SYSTEM MODEL AND METHOD**

  This chapter presents the system design and details the experimental methods used in the study.
- **CHAPTER IV: EXPERIMENTAL RESULTS AND ANALYSIS**

  This chapter covers the discussion of experimental results and their corresponding analysis.
- **CHAPTER V: CONCLUSION**

  This chapter summarizes the essential findings and conclusions drawn from the research.