Transformer Augmented YOLOv8 For Small Drone Detection Using Deep Learning

Rubi Bency .M, Jasmi Priya. J, Dharshana .VS, Ebi T Prabha,
Department of Information Technology, Bethlahem Institute of Engineering, Karungal
AP/IT, Bethlahem Institute of Engineering, Karungal

Abstract: **The increasing misuse of drones poses significant safety and security risks, including illegal transportation of prohibited goods, interference with manned aircraft, and threats to public safety. This has raised concerns about the increased use of unmanned aerial vehicles (UAVs) due to their small size. Addressing these concerns has sparked significant research into developing effective drone detection systems. Deep learning, especially YOLO, is known as a lightweight model that offers real-time detection capabilities. Attention mechanisms have proven effective in many studies for detecting objects. This research focused on optimizing the YOLOv8n-based model by incorporating the Attention Module into the neck and improving the detection head by adding a tiny detection head, making the model work efficiently in detecting objects of tiny size. To obtain the most effective model, multiple training sets have been experimented with involving different types of attention modules, such as the Convolutional Block Attention Module (CBAM), ResBlock CBAM, Global Attention Mechanism (GAM), and Efficient Channel Attention (ECA). Therefore, based on the results, YOLOv8n + ResCBAM + high-resolution detection head, called P2-YOLOv8n-ResCBAM significantly improves the mean Average Accuracy (mAP) from 90.3% to 92.6%. Although the increased model complexity reduced frames per second (fps) from 263 to 166, the detection speed remains suitable for real-time applications. The proposed model effectively distinguishes drones from birds and recognizes them at long distances, demonstrating its potential for enhancing aerial surveillance and security measures.**

I.INTRODUCTION

The rapid spread of drones has led to significant progress in multiple industries, such as delivery, agriculture, and surveillance. Nevertheless, the rise in drone utilization has resulted in heightened apprehensions regarding security and privacy, including the illicit conveyance of prohibited items, disruption of manned aircraft operations, and jeopardizing public well-being. To tackle these challenges, efficient drone detection systems must be used to accurately differentiate drones from other objects in real-time. Drones come in many sizes, from small to large, and are often used in many industries such as monitoring, transportation, communication, and photography [1], [2], [3]. Hence, the proliferation of drones proves the advantages of improving our daily lives [4]. Nevertheless, while drones provide numerous advantages to society, their misuse can pose significant risks to safety, privacy, and security [5], [6]. Threats include privacy invasion, target attacks, breach of the No-Fly Zone (NFZ), and illegal transportation for smuggling, such as explosive things or drugs [3], [5], [7], [8], [9], [10], [11], [12]. The increase in the use of drones has sparked public concerns, and if this trend continues unabated, we may face a future where the sky is drowned by drones [3]. Therefore, implementing a drone detection system is an important step in reducing and dealing with this issue. This has increased researchers' awareness of the importance of developing an effective drone detection system. Various techniques have been introduced in the development of this system. Compared to traditional methods, deep learning offers superior capabilities in automatically extracting and learning target features directly from data [13]. Deep learning is a branch of artificial intelligence (AI), using neural networks to process data. Through machine learning, these networks can be trained on vast data sets, allowing them to learn and recognize patterns autonomously. In essence, AI systems can emulate the functioning of the human brain by predicting outcomes based on observed patterns. Due to the advanced technology, its increasing popularity can be attributed to the accessibility of training data, advanced hardware, and computational resources [14], which have significantly expanded the use of deep learning techniques. As a result of these technological advances, the use of deep learning is increasing in various industries, especially in object detection [15]. Various techniques are available within the realm of Convolutional Neural Networks

(CNNs). Object detection stands out as a superior choice over conventional radar and infrared in developing drone detection system [6], [16]. In general, object detection involves two main tasks: localization and classification. This task aims to determine the exact location of a target object in an image or video and identify the category of the object. Object detection has two main approaches: one-stage and two-stage detectors. One-stage detectors, such as You Only Look Once (YOLO) [17] and Single Shot MultiBox Detector (SSD) [18], directly predict bounding boxes and class labels at the same stage. In contrast, two-stage detectors, such as Region-based Convolutional Neural Networks (R-CNN) [19], Fast R-CNN [20] and Faster R-CNN [21], involves two stages to identify the potential area and classify the targeted objects within this area. A study conducted by [22] aimed to determine the optimal model between Faster R-CNN, YOLO, and SSD to detect drones in various environments, focusing speed and accuracy. The results demonstrates that although SSD better in detection ability, Faster R-CNN and YOLO exhibit superior recognition abilities. However, according to [23], among various algorithms under object detection, YOLO offers a balanced combination of speed and accuracy, making it as a fast and reliable detection model. YOLO was designed expressly to overcome problems involving speed of inference while conserving competitive accuracy [6]. This is achieved by simultaneously performing bounding box determination and classification in the same stage. Researchers have continued to improve YOLO since its launch in 2015, leading to multiple versions. Figure 1 shows a timeline for the many variants of YOLO. Figure 2 gives an overview of its general functionality. When choosing an efficient model to create a real-time drone detection system, YOLO-based models have performed well in many studies. As evidence, many studies have proven its effectiveness by demonstrating good results in terms of accuracy and speed. The YOLOv4 model was chosen for developing a drone detection system in a study by [9]. Using the transfer learning method, this model has outperformed Faster R-CNN in terms of mean Average Precision (mAP) after tuning the model. A study by [24] that has trained a fine-tuned YOLOv5 model, has achieved the highest precision, 97.4 % and at the same time surpassed YOLOv3, YOLOv4 and Mask R-CNN. In addition, this model also demonstrated a good performance in detecting small drones. To detect small objects and balance between accuracy and speed, [25] has modified the YOLOv8m-based model by adding a P2 Layer and Multi-Scale Image Fusion (MSIF). The highest frame per second (fps), 45.7 fps, shows that this method can detect drones very fast, even if they are small objects. The capability of the attention mechanism to improve the model performance during training is proven by many studies, such as [6], [26], [27], [28], [29], [30], [31], and [32]. Selective hearing refers to the cognitive process in which individuals prioritize and concentrate on significant auditory stimuli amidst a distracting or noisy environment, rather than attempting to process all auditory information simultaneously. The attention module in deep learning operates on the same principle, enabling the model to concentrate on significant elements while disregarding irrelevant information selectively. Due to various elements like edges, textures, and background in the input image, the model faces challenges in accurately identifying the specific object in the image. By incorporating an attention module into the original architecture, the model can effectively concentrate on capturing the specific details of the targeted object, resulting in improved performance. Despite technological progress, the task of detecting small drones remains difficult because of their compact size and the intricate process of differentiating them from similar entities such as birds. The objective of this study is to improve the performance of the YOLOv8 model by integrating different attention modules, thereby increasing its precision in detecting small drones. This paper proposes an effective method to develop drone detection that can distinguish between birds and drones and detect them even at long distances

II. RELATED WORKS

Builds a new dataset by gathering drones and birds of a small size from numerous available datasets.Employs four different attention mechanisms to the neck part of the YOLOv8n model where it involves Convolutional Block Attention Module (CBAM), Global Attention Mechanism (GAM), Efficient Channel Attention (ECA), and ResBlock CBAM (ResCBAM) during training. Adds high-resolution head to the head part, where it increases the model capability in detecting small targets. Tunes the hyperparameters during training. Ablation tests are carried out for every attention module, with and without a high-resolution head, utilizing different hyperparameter sets. The goal of these experiments is to find the best model, and

ResCBAM + high-resolution head + tune hyperparameters achieve the best mAP.

This paper is organized as follows: Section II presents and explains the original architecture of YOLOv8 and the proposed version. Section III displays the training platform, including the software and hardware requirements and the training setup. This section also describes the dataset used and how the model is trained. Several experiments were conducted to compare each of the models with other YOLO versions. Next, section IV analyses the results based on the experiments. Section V presents the detection result based on the P2-YOLOv8n-ResCBAM model, and the last section concludes the overall work.

III. METHODOLOGY

Visual Drone Detection System is a system designed to identify and categorize objects of interest, including drones, by visual means. The presence of the drones is then recognized by extracting their characteristics from the captured image. The optimized drone detection model will be built on the foundation of the YOLOv8 model. A. YOLOv8 MODEL YOLOv8 offers five sizes, and in this research, the smallest model, YOLOv8n, is selected. Three main parts can be divided to represent YOLOv8 architecture: backbone, neck, and head. The backbone is responsible for extracting meaningful features from input images at various scales, the neck is known as multi-feature fusion, where all extracted features from different layers will be combined to get meaningful information, and the head works to make predictions. In the development of YOLOv8, three important elements can be highlighted based on [33]: 1. New convolution, C2f module is replaced C3 block as main building block in YOLOv8. To build C2f, the concept of ELAN (Efficient Layer Aggregation Network) [34] is used to improve resonance speed [35]. Unlike C3, which only uses the last bottleneck output, all bottleneck outputs in C2f will be combined. The idea is comparable to the ResNet Block [36]. 2. Anchor-free detection is utilized instead of predicting using an offset bounding box known as an anchor box like other models. This means it will predict directly from the center of an object. This innovation has reduced the number of overlapped prediction boxes, which can accelerate the Non-Maximum Suppression (NMS). 3. Closing the mosaic augmentation for the last 10 epochs during training. During training, by applying mosaic augmentation, the model can learn objects in different locations as four images will be gathered in one image together. However, this augmentation can somehow decrease performance, and it is believed that turningoff this augmentation for the last 10 epochs can prevent the deterioration. Several innovations in the development of YOLOv8 have contributed to an increase in accuracy and beat YOLOv5 and YOLOv7 when tested using dataset Microsoft COCO and Roboflow 100.

IV. DATA PREPARATION

1) HARDWARE AND SOFTWARE REQUIREMENTS

This research has been trained using Intel i9-14900K with 64GB memory. It allows multitasking and efficient handling of big and complex data. A large storage device, a 1.8 TB SSD drive, and a high-powered GPU, NVIDIA GeForce RTX 4090, are used to accelerate the training process. Several software specifications are required to perform this research, and the details are displayed in Table

2) TRAINING SETUP

The training setup includes hyperparameters for model training are listed in Table 2. B. DATASET CONSTRUCTION This research aims to develop a drone detective system that is able to differentiate between drones and drone-like objects, such as birds, and track them even at long distances Regardless of the actual distance of the object from the source, its small size in the frame, resulting in small pixels, can represent how far the drone or bird is from the camera. Therefore, it is crucial to provide a dataset that meets those criteria to allow the model to learn effectively. Hence, a new dataset, BirDrone [41] was prepared by collecting images of small drones, including multirotor types, such as quadcopters, hexacopters, and octocopters, as well as birds from different datasets [42], [43], [44], [45], [46], [47]. We have included images with multiple drones or birds in one image for model training. The YOLO framework itself is designed to detect multiple objects in one frame. The proof of detection will be shown in Section V. This dataset also includes different types of backgrounds and lighting. Fig. 6 and 7 show examples of raw images in our dataset. Before proceeding to training, the dataset needed to be annotated first, and it was done manually using Roboflow. The smallest annotation bounding box, which represents the size of the targeted object, is 7×14 pixels, and the largest one is 65 × 182 pixels. Then, the dataset went through several pre-processing methods, such as auto orient, which can standardize overall orientation, improve overall analysis, and be suitable for real-time applications. Lastly, auto-adjust contrast makes the details about birds and drones easier to see. Next, a data augmentation approach is used for the model to identify drones and birds in various scenarios. For each image, geometric transformations such as rotation and exposure are used. By displaying multiple perspectives, rotation can increase model robustness and help avoid overfitting. Applying exposure to the images gives more variability to the dataset regarding lighting and environment. After that, the total of images 2970 was divided into 80% for training and 20% for validation.

MODIFICATION ON THE PRE-TRAINED RANDOM FOREST ALGORITHM FOR HEART DISEASE PREDICTION

Several training sessions were carried out in this section using a designated dataset to verify various aspects of the proposed model. Initially, the effectiveness of the supplementary detection head and the P2-YOLOv8n model, specifically developed for identifying minuscule entities, was evaluated. Furthermore, an assessment was conducted to determine the effect of incorporating attention modules into the YOLOv8- based model. Furthermore, an analysis was conducted to evaluate the efficacy of optimizing hyperparameters. The comparison was evaluated based on precision, recall, and mAP, and the formula. depicts a training sample utilized in these experiments, demonstrating the model's ability to manage diverse training scenarios and configurations effectively. Table 4 has demonstrated the training results of several YOLOv8nbased models with several attention modules but using default hyperparameters. Table 5 presents the training results of the YOLOv8n-based model but with a tuning hyperparameter for the optimizer, which SGD [48] is used, and the value of weight decay was set to 0.00015 as a recommendation from Ultralytics [49]. Table 6 shows the training results when using 0.73375 for momentum value as a recommendation from Ultralytics [49], 0.00015 for weight decay and SGD for the optimizer. Table 7 displays the training results of several YOLOv8n-based models when tuning the hyperparameters, with the momentum value set to 0.94, a slight increment from the previous table. Also, weight decay was set lower than before, at 0.00012. The momentum value was set to 0.942, a slight increase from Table 5 and Table 6, which used 0.0005 for weight decay, and the rest of the hyperparameters were the same as Table 5. Finally, Table 10 shows the training results for the proposed model for all classes, including both drone and bird, drone only, and bird only.

V. THE PROPOSED SYSYTEM

Backbone YOLOv8 backbone for feature extraction. Transformer Module Integrated after backbone to enhance contextual information, especially for small targets. Neck and Head: YOLOv8's PAN-FPN or BiFPN for feature fusion and object detection. Training Dataset: Aerial drone datasets (e.g., Anti-UAV, Drone-vs-Bird). Evaluation Metrics mAP@0.5, precision, recall, and F1 score.

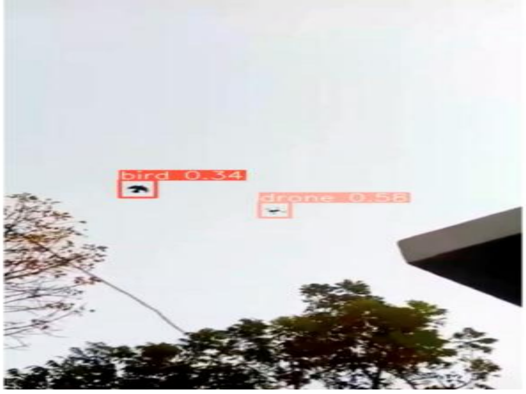
1. Input Image: Aerial surveillance images/video frames.
2. YOLOv8 Backbone: Extracts initial spatial features.
3. Transformer Encoder Module: Captures global context. Refines features by modeling long-range dependencies.
4. YOLOv8 Neck: Merges features at multiple scales.
5. YOLOv8 Head: Predicts bounding boxes and class probabilities.

RESULTS

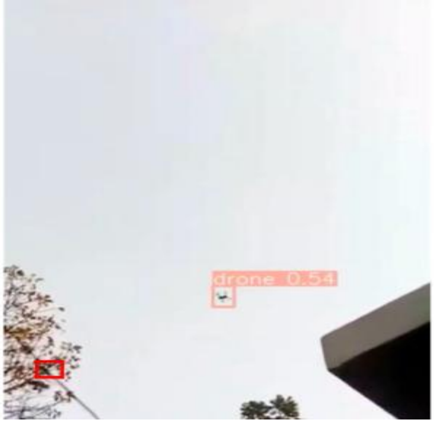
INPUT

OUTPUT:

MODEL DEPLOYEMENT FOOTAGE



BIRD DETECTION FAILURE



P2-YOLOv8n-ResCBAM MODEL

VI. CONCLUSION

In this conclusion, Detecting a drone with dynamic movements, small size, and even a shape similar to a bird is indeed challenging. Therefore, building an accurate model that can detect in real-time is crucial. However, speed and accuracy always have a trade-off between them. By using YOLOv8n as a base model, this research proposed integrating an attention module and adding an extra high-resolution detection head. To support this proposed model to reduce false detection, the new dataset has been created to provide effective learning for the model to learn how to differentiate between drones and birds as well as recognize them even at long distances. Powerful hardware is utilized to ensure the inference speed is aligned with real-time detection. Fine-tuning the hyperparameters is also one of the methods used to optimize training performance, which can lead to better detection. Based on training results, the P2-YOLOv8n-ResCBAM model has demonstrated improvement in mAP, which is from 90.3% to 92.6%, showing a 2.3% increment. However, due to the noticeable increment in model parameters, it can be noticed that the fps is decreased from base model during deployment, which is from 263 fps to 166 fps, but the fps achieved remains suitable in real-time detection. It is also believed to be why the inference speed in fps decreased. Apart from that, the model deployment result also portrayed a good result, where the model was able to differentiate between drones and birds even at long distances by using video and images as input. However, the model struggles to detect objects when they overlap with other objects in the same frame. Firstly, considering the model's complexity, the inference speed is fast due to the powerful hardware used during training and deployment. Therefore, the performance may differ depending on the type of hardware used, especially if low-end hardware is used. While the model may not be ideal for implementation on low-end hardware, the complexity of the model could be justified by its capability to detect a wide range of targets, especially tiny ones, and differentiate between drones and birds. Therefore, the decision to use the proposed model should be based on weighing these potential benefits against the hardware requirements. Next, they have built advanced drone technology that mimics birds' looks and behavior, such as the eagles. Although the proposed model is being trained to differentiate between drones and birds, the model may not detect this type of technology as the detection is only based on visuals. Therefore, this research could further explore the application as there is room for further improvement, such as model size reduction, and optimize this method to make it suitable for industry needs.