Sarthak Saini

A Lightweight ATTENTION- GUIDED BIFPN-YOLO11 model FOR DRONE DETECTION

Thesis Mtech

Document Details

Submission ID

trn:oid:::27535:140830714

Submission Date

May 29, 2026, 1:12 AM GMT+5:30

Download Date

May 29, 2026, 1:17 AM GMT+5:30

File Name

24ISY111_Sarthak Saini-5-45.pdf

File Size

1.9 MB

41 Pages

11,320 Words

69,382 Characters

8% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Exclusions

- 1 Excluded Source

Match Groups

- **80** Not Cited or Quoted 7%
  Matches with neither in-text citation nor quotation marks
- **0** Missing Quotations 0%
  Matches that are still very similar to source material
- **12** Missing Citation 1%
  Matches that have quotation marks, but no in-text citation
- **0** Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

4% Internet sources
5% Publications
6% Submitted works (Student Papers)

Match Groups

- **80** Not Cited or Quoted 7%
  Matches with neither in-text citation nor quotation marks
- **0** Missing Quotations 0%
  Matches that are still very similar to source material
- **12** Missing Citation 1%
  Matches that have quotation marks, but no in-text citation
- **0** Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

4% Internet sources
5% Publications
6% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Internet	www.mdpi.com	<1%
2	Internet	arxiv.org	<1%
3	Internet	ejournal.ikado.ac.id	<1%
4	Internet	dl.lib.mrt.ac.lk	<1%
5	Student papers	York St John University on 2025-08-18	<1%
6	Student papers	University of Westminster on 2026-05-24	<1%
7	Internet	nrl.northumbria.ac.uk	<1%
8	Publication	Xiu Chen, Yujie Li, Yoshihisa Nakatoh. "Pyramid attention object detection netwo...	<1%
9	Publication	M. Raja, Satya Subrahmanyam, R. Raja Subramanian, J. Karthikeyan. "Adaptive Te...	<1%
10	Publication	Wentian Zhao, Xinxiao Wu, Jiebo Luo. "Cross-domain Image Captioning via Cross-...	<1%

11 Internet

dspace.daffodilvarsity.edu.bd:8080 <1%

12 Student papers

Liverpool John Moores University on 2024-08-30 <1%

13 Student papers

University of Hong Kong on 2026-04-20 <1%

14 Student papers

University of Leeds on 2026-05-06 <1%

15 Internet

harvest.usask.ca <1%

16 Publication

Muhammad Haris Kaka Khel, Kushsairy Kadir, Sheroz Khan, Mnmm Noor, Haidaw... <1%

17 Publication

Xinhua Mao, Zhimin Chen, Mengqi Zhang, Jinwei Sun, Chengshan Xu. "Particle Vel... <1%

18 Internet

docplayer.net <1%

19 Internet

dspace.iiti.ac.in <1%

20 Publication

B. G. Nagaraja, S. Kannadhasan. "Information and Communication Systems", CRC... <1%

21 Publication

Haomin Li, Huanzun Zhang, Wenke Zang. "StripSurface-YOLO: An Enhanced Yolov... <1%

22 Student papers

University of Hertfordshire on 2026-04-26 <1%

23 Publication

Lili Wang, Lei Guo, Haiyan Li, Bingbing He, Jundong Yang, Yaqun Huang. "Enhanc... <1%

24 Internet

theses.hal.science <1%

25 Internet
afribary.com <1%

26 Internet
aichutan.github.io <1%

27 Internet
digiresearch.vut.ac.za <1%

28 Internet
ualberta.scholaris.ca <1%

29 Internet
documents1.worldbank.org <1%

30 Internet
eprints.soton.ac.uk <1%

31 Internet
ir.nctu.edu.tw <1%

32 Publication
"Pattern Recognition. ICPR International Workshops and Challenges", Springer Sc... <1%

33 Publication
Allan J. Kember, Hafsa Zia, Praniya Elangainesan, Min-En Hsieh et al. "Transitionin... <1%

34 Student papers
Quest International University on 2026-03-17 <1%

35 Student papers
University of Huddersfield on 2022-04-18 <1%

36 Student papers
University of Newcastle upon Tyne on 2025-08-27 <1%

37 Student papers
University of Sheffield on 2026-05-13 <1%

38 Internet
avesis.erciyes.edu.tr <1%

39 Internet
eprint.iitd.ac.in <1%

40 Internet
repository.essex.ac.uk <1%

41 Internet
spectrum.library.concordia.ca <1%

42 Internet
www.aimspress.com <1%

43 Publication
Alshamsi, Omar. "Intelligent Monitoring and Alerting System Using Some Deep L... <1%

44 Student papers
Binus University International on 2024-12-16 <1%

45 Publication
Guofan Long, Wei Song, Xiangchun Liu, Ziyao Fang, Jinqi An, Kun Liu, Yaqin Huan... <1%

46 Student papers
Higher Education Commission Pakistan on 2025-01-08 <1%

47 Student papers
UCL on 2025-09-08 <1%

48 Student papers
University of Sydney on 2024-05-27 <1%

49 Internet
www.ijraset.com <1%

50 Publication
Arvind Dagur, Sohit Agarwal, Dhirendra Kumar Shukla, Shabir Ali, Sandhya Sharm... <1%

51 Student papers
Assam University, Silchar on 2026-05-20 <1%

52 Publication
J. Anitha Menon, Benita Veronica, Ben Milbourn, Mithra S. Gnana Sanga. "Sustain... <1%

53 Publication

Jaskaran Singh, Meenu Gupta, Rakesh Kumar. "Smart Technologies and Intelligen... <1%

54 Student papers

Kingston University on 2025-09-05 <1%

55 Student papers

Macao Polytechnic Institute on 2026-04-15 <1%

56 Student papers

Nazarbayev University on 2022-04-28 <1%

57 Student papers

SASTRA University on 2026-03-16 <1%

58 Student papers

University of Greenwich on 2025-04-30 <1%

59 Student papers

University of Greenwich on 2025-12-17 <1%

60 Student papers

University of Sydney on 2021-10-22 <1%

61 Publication

Ziran Ye, Yongyong Fu, Muye Gan, Jinsong Deng, Alexis Comber, Ke Wang. "Buildi... <1%

62 Internet

idoc.tips <1%

63 Publication

"Pattern Recognition and Computer Vision", Springer Science and Business Media... <1%

64 Publication

"Advanced Intelligent Computing Technology and Applications", Springer Science... <1%

65 Publication

"Image and Graphics", Springer Science and Business Media LLC, 2021 <1%

66 Publication

"Pattern Recognition and Machine Intelligence", Springer Science and Business M... <1%

67 Student papers

Associatie K.U.Leuven on 2026-05-22 <1%

68 Publication

Bal S. Virdee, Tanweer Ali, Jaume Anguera, Suman Lata Tripathy. "Connecting Int... <1%

69 Student papers

National Taiwan University on 2023-08-08 <1%

70 Publication

Paolo Ferro, Harinadh Vemanaboina, Chander Prakash. "Computational Techniqu... <1%

71 Student papers

University of Nottingham on 2022-12-15 <1%

A Lightweight ATTENTION- GUIDED BIFPN-YOLO11 model FOR DRONE DETECTION

Sarthak Saini

ABSTRACT

With the fast increase in the number of applications that UAVs perform in areas of surveillance, security, disaster management, and commercial operations, there has developed a very significant requirement for efficient and accurate real-time drone detection systems. Nevertheless, the process of detecting UAVs poses great challenges owing to factors such as small object size in image frames, changes in illumination levels, complexity of scenes and motion variations. While YOLO-based deep learning methods achieve real-time capabilities, they still struggle in detecting small UAVs and coping with complex environments.

This thesis presents a state-of-the-art UAV object detector model based on YOLO11. This architecture is enhanced using two additional models – the Convolutional Block Attention Module (CBAM) and the Bi - directional Feature Pyramid Network (BiFPN). The purpose is to increase the ability of this model to detect smaller objects while maintaining efficiency and fast operation. The first enhancement allows improving the detection of key features of the drone objects both spatially and through channel attention.

The other modification makes it possible for the model to improve its multi-level feature representations since it facilitates two-way fusion of the features from different pyramid levels, which helps in localization of the drones at different levels. The models were trained and tested using the open - source dataset YOLO Drone Detection Dataset, available at Kaggle, that contains numerous examples of indoor, outdoor, distant, and outline drones having 347 samples for validation along with 369 labeled drones. Both YOLO11n and YOLO11s models were experimented with extensively using the same training set-up. For performance evaluation, the standard object detection parameters such as Precision , Recall , mAP@50 , mAP@50-95 , F1-Score and Frames Per Second (FPS) were used.

The experimental results show that YOLO11s + CBAM + BiFPN obtained a Precision of **94.8%**, Recall of **93.7%**, mAP@50 of **96.1%**, mAP@50-95 of **71.8%** and F1-Score of **94.2%**. These correspond to gains of +2.4%, +5.6%, +2.5%, and +3.3% respectively compared to the plain YOLO11s baseline, while still running in real-time at **75 FPS** on an NVIDIA Tesla T4 GPU. Further analysis of qualitative confidence demonstrated that YOLO11s + CBAM + BiFPN had an average detection confidence of 16.4% higher than YOLO11n + CBAM + BiFPN on the same input frames, indicating the benefit of deeper feature extraction for small drone targets.

68
63

Collectively, the results indicate that the combination of lightweight attention modules and advanced multi - scale features and fusion can significantly improve the accuracy and robustness of UAV detection. The devised approach exhibited strong robustness under difficult circumstances, including extremely small and faraway drones, reflection, occlusions, and contrastless images. This is a viable and scalable solution to many practical problems such as drone tracking and surveillance.

CONTENTS

CANDIDATE'S DECLARATION ..... **II**
ACKNOWLEDGEMENT ..... **III**
CERTIFICATE BY THE SUPERVISOR ..... **IV**
ABSTRACT ..... **V**
CONTENTS ..... **VII**
LIST OF FIGURES ..... **X**
LIST OF TABLES ..... **XI**
LIST OF SYMBOLS, ABBREVIATIONS AND NOMENCLATURE ..... **XII**
CHAPTER 1 INTRODUCTION **1**
1.1 BACKGROUND AND MOTIVATION 1
1.2 REAL - TIME OBJECT DETECTION AND YOLO FRAME-WORK 1
1.3 ROLE OF ATTENTION MECHANISMS AND MULTI-SCALE FUSION 2
1.4 PROBLEM STATEMENT 2
1.5 OBJECTIVES OF THE RESEARCH 3
1.6 PROPOSED METHOD 3
1.7 RESEARCH CONTRIBUTIONS 4
1.8 ORGANIZATION OF THE THESIS 4
CHAPTER 2 LITERATURE REVIEW **5**
2.1 INTRODUCTION TO THE REVIEW 5
2.2 TRADITIONAL UAV DETECTION METHODS 5
2.3 EVOLUTION OF OBJECT DETECTION: FROM R-CNN TO YOLO 6
2.4 ATTENTION MECHANISMS IN CONVOLUTIONAL NEURAL NETWORKS 6
2.5 MULTI-SCALE FEATURE FUSION ARCHITECTURES 8
2.6 SMALL OBJECT AND AERIAL OBJECT DETECTION 8
2.7 UAV DETECTION USING DEEP LEARNING 9
2.8 SUMMARY AND RESEARCH GAP 10
CHAPTER 3 PROPOSED METHODOLOGY **12**
3.1 OVERVIEW OF PROPOSED FRAMEWORK 12
3.2 PIPELINE OF IMPLEMENTATION 13

3.2.1 Dataset Description 13
3.3 DATA - PREPROCESSING AND DATA – AUGMENTATION 13
3.3.1 Preprocessing 13
3.3.2 Data Augmentation Scheme 14
3.4 YOLO11S BACKBONE DESIGN 15
3.5 SELECTIVE CBAM 16
3.5.1 Motivation 16
3.5.2 Placement Strategy 16
3.5.3 CBAM Architecture 16
5.1 LIGHTWEIGHT BIFPN INTEGRATION 17
5.1.1 Motivation 17
5.1.2 Lightweight Design 18
5.1.3 Bidirectional Feature Fusion 18
5.2 TRAINING METHODOLOGY 19
5.2.1 Hyperparameter Configuration 19
5.2.2 Loss Functions 19
5.3 SUMMARY 19
CHAPTER 6 RESULTS AND DISCUSSIONS 21
6.1 OVERVIEW 21
6.2 EXPERIMENTAL ENVIRONMENT 21
6.2.1 Hardware and Software Configuration 21
6.3 TRAINING CONFIGURATION 21
6.3.1 Dataset and Augmentation 21
6.4 MODELS EVALUATED 21
6.5 EVALUATION METRICS 22
6.6 QUANTITATIVE RESULTS 22
6.6.1 Overall Performance Comparison 22
6.7 ANALYSIS BY ARCHITECTURE SCALE 23
6.7.1 YOLO11n: Modest but Consistent Gains 23
6.7.2 YOLO11s: Substantial Gains Across All Metrics 24
6.8 PRECISION–RECALL ANALYSIS 25
6.9 QUALITATIVE DETECTION RESULTS 25
6.10 DISCUSSION 27
6.11 LIMITATIONS 28
6.12 SUMMARY 28

6 **CHAPTER 7 CONCLUSION AND FUTURE SCOPE 29**
7.1 OVERVIEW 29
7.2 RESEARCH SUMMARY 29
7.3 MAJOR FINDINGS 29
7.4 CONTRIBUTIONS OF RESEARCH 30
22 7.5 PRACTICAL APPLICATIONS 31
7.6 LIMITATIONS OF THE CURRENT WORK 31
7.7 FUTURE SCOPE 32
7.8 FINAL CONCLUSION 33
CHAPTER 8 REFERENCES 34

25

LIST OF FIGURES

Figure 3.1: Overview of the proposed pipeline.....12
Figure 3.2: Proposed backbone design of YOLOs + CBAM + BiFPN 15
Figure 4.1: All-metric performance comparison across evaluated architectures 23
Figure 4.2: Metric-wise improvements (%) after CBAM and BiFPN integration.....24
Figure 4.3: Precision–Recall distribution with iso-F1 contours 25
Figure 4.4: Qualitative drone detection results across four surveillance scenarios 26
Figure 4.5: Detection confidence comparison: YOLO11n vs. YOLO11s on the same frame 27

39

39

LIST OF TABLES

Table 2.1: Comparative summary of representative UAV detection works 10
Table 3.1: List of symbols used in above equations.....17
Table 3.2: Table of Hyperparameters 19
Table 4.1: Performance Comparison of Baseline and Proposed Models 22
Table 4.2: Metric-Wise Improvements After CBAM and BiFPN Integration 24

LIST OF SYMBOLS, ABBREVIATIONS AND NOMENCLATURE

UAV	Unmanned Aerial Vehicle
YOLO	You Only Look Once
CBAM	Convolutional Block Attention Module
BiFPN	Bi-directional Feature Pyramid Network
FPN	Feature Pyramid Network
PANet	Path Aggregation Network
CNN	Convolutional Neural Network
mAP	Mean Average Precision
IoU	Intersection over Union
FPS	Frames Per Second
TP	True Positive
FP	False Positive
FN	False Negative
SGD	Stochastic Gradient Descent
AdamW	Adaptive Moment Estimation with Weight Decay
ReLU	Rectified Linear Unit
SiLU	Sigmoid Linear Unit
DFL	Distribution Focal Loss
CIoU	Complete Intersection over Union
BCE	Binary Cross-Entropy
GPU	Graphics Processing Unit
RF	Radio Frequency
SENet	Squeeze-and-Excitation Network
BAM	Bottleneck Attention Module
ViT	Vision Transformer
F	Input feature map
$M_c(F)$	Channel attention map
$M_s(F)$	Spatial attention map
w_i	Learnable BiFPN fusion weight for feature level i
ϵ	Small numerical stability constant (ϵ= 10−4)
$\sigma(\cdot)$	Sigmoid activation function
C_{out}	Output channel dimension of a convolutional layer
$P3, P4, P5$	BiFPN feature levels (small, medium, large drones)

41

CHAPTER 1

INTRODUCTION

1.1 BACKGROUND AND MOTIVATION

Today's drones are equipped to carry out real-time imagery, navigation, cargo transport, and intelligent surveillance. As a result, they can be used for various sophisticated operations. Unfortunately, the rise in the number of UAVs has also led to a number of security and safety risks. Drones flying without authorization around sensitive facilities, such as airports, military bases, border regions, prisons, and vital industries, have become common lately. The use of drones in this way has given rise to issues such as espionage, smuggling, terrorism, privacy invasion, and unlawful flight over important places [1,2].

In addition, it has become very hard to spot these aircraft because of their small size, fast speed, and ability to fly autonomously. For that reason, it is now essential to develop effective methods to detect such drones.

The old traditional methods used to identify UAVs such as the use of radars, acoustic signals, and RF signals have been characterized with various limitations when tried to detect small and low altitude unmanned drones. Small drones usually have very low radar signatures, weak acoustic signatures, and unpredictable and varied radio frequencies, which makes them undetectable by conventional methods [3].

As a solution to these problems, vision-based deep learning algorithms have become one of the most prospective options for drone detection [4]. The ability of deep learning algorithms to discover discriminative visual features on their own makes drone detection feasible even in difficult environments. The adaptability of CNNs to various backgrounds, different illumination levels, motion blur, and scale variations surpasses that of conventional algorithms based on hand-crafted features [5].

1.2 REAL - TIME OBJECT DETECTION AND YOLO FRAME-WORK

44

Recently, YOLO (You Only Look Once), in the object detection model [6] have become one of the most popular options out of all available object detectors. What sets the YOLO algorithm apart from other two - stage detectors like Faster R-CNN [7] is that it does not generate any region proposal prior to classification; it is an algorithm which detects and classifies at the same time in one forward pass of the image through the network.

The latest generation of YOLO algorithm, the YOLO11, [8] increases the processing speed by means of an upgraded backbone design and improved feature extraction process, in addition to representing images at multiple scales.

- **Small Object Size:** The drone is usually visible in just a few pixels in the images, particularly for distant monitoring tasks [9].
- **Complex Background:** The sky, buildings, trees, clouds, and reflective surfaces cause background clutter leading to misclassification.
- **Lighting Variations:** The visibility of the drones is highly dependent on factors such as haze, fog, low light environment, glare, and backlit by the setting sun.
- **Motion and Blur:** Quick movement of the drone and camera motion may cause distortion, resulting in reduced detectability. [10].
- **Visual Similarity:** Drones may look like birds, planes, or other objects far away from the viewer. [11].

The solution to such problems involves the design of sophisticated detection mechanisms that not only have the ability to learn discriminative features but can also concentrate on regions of interest from a drone perspective at various levels of spatial detail.

1.3 ROLE OF ATTENTION MECHANISMS AND MULTI-SCALE FUSION

The new developments in the deep learning have shown that the addition of attention blocks increases the ability of the representation by convolutional neural networks [12]. Attention blocks allow CNNs to give importance to significant spatial and channel features while de-emphasizing unwanted backgrounds [13].

Some of the best lightweight attention modules include Convolutional Block Attention Modules (CBAM) , which consists of the channel - attention followed by spatial – attention [14]. These attention blocks help in enhancing the capability of the network which focuses on the features of the drones such as the edge features, shape of the silhouette, structure, and motion features [15]. The low complexity of CBAM makes it very useful for detecting drones in real-time.

However, multi-scale feature fusion has become an essential part of modern object detectors as well. The fact that UAVs can appear at multiple scales based on the height from where they fly and the camera's vantage point means that fusion at various scales becomes essential. The common approaches to performing feature fusion involve the use of Feature Pyramid Networks (FPN) [16], and Path Aggregation Networks (PANet) [17]. FPN and PA-Net improve object detection results via feature fusion of both spatial and semantic feature maps.

The Bi - directional Feature Pyramid Network (Bi-FPN) is another approach that improves upon FPN and PANet by enabling efficient bidirectional propagation of features, along with learning-based feature fusion.

1.4 PROBLEM STATEMENT

Although existing YOLO-based drone detection systems [18, 19, 20] have shown strong real-time performance but their effectiveness often drops in scenarios involving:

- Tiny and distant drones,
- Complex and occlusive backgrounds,
- Poor illumination conditions,
- Reflective environments,
- Low-contrast aerial scenes.

Many existing and old approaches either prioritizes detection accuracy over the cost of computational complexities or solely focuses on lightweight deployment while sacrificing robustness [1]. Furthermore, limited research has explored the combined integration of lightweight attention mechanisms and advanced bidirectional multi-scale fusion within an advance and new YOLO11 framework specifically for drone detection applications.

Therefore, there exists a need of lightweight yet highly accurate drone detection architecture capable of improving small-object sensitivity also at same time preserving real-time inference capability [9].

1.5 OBJECTIVES OF THE RESEARCH

Major objective of the research is to develop an improved UAV detection framework by fusing CBAM attention [14] and BiFPN feature [21] inside the YOLO11 architecture [8]. The below approach also aims to the improvement of the feature extraction and multi-scale feature representation of the drones, thereby increasing the accuracy of the small – drone detection and enhancing the robustness under complex real-world conditions and scenarios.

The major objectives of the research work are:

- To improve small and tiny drone detection performance by using the attention-driven feature refinement and optimization.
- To increase multi-scale feature representation through BiFPN fusion.
- To check and evaluate the ability of our proposed architecture under tough conditions and challenging drone surveillance scenarios.
- To compare the baseline of YOLO11 models with enhanced and improved CBAM + BiFPN variants through standard object detection metrics and scales .
- To preserve the lightweight computation and also to infer the real time real-time inference performance suitable for practical deployment even on small devices.

1.6 PROPOSED METHOD

In the work, we proposed a new drone detection system based on YOLO11 [8] with CBAM attention [14] and BiFPN multi-scale feature fusion [21].

The workflow of this system can be summarized as follows:

- Initially , the **YOLO11 backbone** extract the hierarchical features from input drone images.
- Next, the **CBAM module** applies attention mechanism to refine these features and select salient information related to drone objects.
- **BiFPN structure** is the third part which is the bi - directional feature pyramid network for the feature fusion.
- Finally all features are passed to the **YOLO detection head**.

The models proposed in this work were trained and tested on the public YOLO Drone Detection Dataset [22], which contains a large number of drone images captured under diverse conditions such as indoor environments, outdoor scenes, far-off ranges, silhouettes, and complex backgrounds.

1.7 RESEARCH CONTRIBUTIONS

The main contributions of this thesis can be summarized as follows:

- Integration of CBAM attention - mechanism [14] in the YOLO11 network architecture, to enhance the discrimination of features for UAVs.
- Using BiFPN [21] instead of traditional feature concatenation to improve the small object detection ability.
- Demonstration of improved detection performance in the presence of environmental challenges.
- Design of a lightweight and robust UAV detection system for real-world deployment.

1.8 ORGANIZATION OF THE THESIS

The following chapters of this thesis are structured as follows:

- **Chapter 2** Literature review of object detection, attention models, feature fusion techniques and UAV surveillance.
- **Chapter 3** Provides details about the approach proposed, data collection process, changes in architecture and training methods.
- **Chapter 4** Tells about the experimental work, evaluation criteria and results obtained.
- **Chapter 5** Defines the conclusions of the thesis and discuss about the possible future directions in this field.

CHAPTER 2

LITERATURE REVIEW

2.1 INTRODUCTION TO THE REVIEW

Vision-based UAV detection which uses deep - learning has grown into a well-established research area over the past decade, largely in response to the widespread deployment of commercial drones in airspace that was previously uncontested. This chapter surveys the literature across six thematic areas that directly inform the proposed framework: traditional UAV detection methods , the evolution of real - time object detection from region-based networks to YOLO, the development of the attention - mechanism for convolutional networks , multi-scale feature fusion architectures, the broader challenge of small and aerial object detection, and UAV-specific deep learning detection systems. The chapter closes with a comparative summary and a precise statement of the research gaps that motivate the proposed work.

2.2 TRADITIONAL UAV DETECTION METHODS

Before deep - learning became relevent for this issue, UAV detection systems relied primarily on the radar, acoustic sensing, and radio-frequency (RF) signal based techniques and analysis. Each technique has its own strengths but also have well-documented weakness which modes that limit practical deployment in different aspects.

Radar looks attractive because it operates passively and no active attention is needed to it, day and night, regardless of visual conditions. Also, the bigger issue was with small UAVs, however, is that their radar cross-sections (RCS) are very very low — when comparing to birds or large insects — making reliable distinction from background clutter a constant issue [3]. Acoustic detection by-passes this and focuses majorly on high-frequency rotor signatures of the multi-rotor platforms, but its effective range is restricted to a few hundred metres under stable environment and degrades exponentially in the presence of environment disturbance, traffic noise, or urban interference [11].

RF-based systems also monitors the communication link between a drone and its operator to identify the presence and, in few cases, also location. This works good for drones operating on certain standard frequency bands majorly the consumer drones, but these autonomous platforms running the pre-programmed missions with their targets fixed without active ground links are completely invisible to RF analysis [15]. Beyond their individual restraints, all three techniques needs a dedicated sensor installations in them which increases the cost to procure, maintain, and scale it across to large perimeters.

Visual detection which uses the standard RGB cameras presents a far more pocket friendly and passive option. The trade-offs are different in different modals: cameras are little sensitive to lighting and weather conditions, and also the combination of small target size, blurry backgrounds, and rapid motions brings a new demanding perception problem.

These attributes are precisely which motivated the shift to learning-based visual detection models.

2.3 EVOLUTION OF OBJECT DETECTION: FROM R-CNN TO YOLO

The new and modern history of the deep learning - based object detection techniques begins with some region - based convolutional neural network (R-CNN) family. The old school R-CNN also used selective search to produce candidate regions, then classified each cropped region independently with a CNN. This two-stage design was accurate and precise but was slow, with evaluation time measuring in tens of seconds per image.

Fast R-CNN [23] improved throughput significantly by adding the CNN to the full image once and then projecting and transfering the region proposals onto the resulting feature map, hence eliminating redundant and ambiguous computation overhead. **Faster R-CNN** [7] went more further by exchanging the external selective search with some learned Region Proposal Network (RPN) that shared convolutional features with the detection - head of the model. Together, all these enhancements brought two-stage detection within reach of the near-real-time operation performance on desktop GPUs, though they are still not that fast enough for edge-deployed surveillance and were very far from a finished modal.

The landscape changed again with **YOLO (You Only Look Once)** [6], which is more to an accurate detection as to a direct regression problem. A single forward - pass on the full image parallely predicts the bounding - box coordinates and the class probability for all the locations over a fixed spatial grid, eradicating the proposal-classification distinction as a whole. The speed advantage over two-stage methods was drastic, and YOLO in a few time became a preferred choice wherever slowness was a big limitation. Newer versions changed the design completely: YOLOv4 [18] also brings CSP backbones, mosaic augmentation, and CIoU loss; YOLOv7 [19] proposed a trainable bag-of-freebies which improved accuracy but not increasing evaluation cost; and YOLOv9 [20] added more programmable gradient information to address information loss through the deep network layers.

The most recent version, **YOLO11** [8], further optimizes the backbone with more depth-wise separable convolutions and also improved C2f modules, and introduces many more capable multi-scale detection heads. These changes and evolvement produces a particularly favourable accuracy-speed tradeoff which makes it suitable for real-time UAV surveillance.

2.4 ATTENTION MECHANISMS IN CONVOLUTIONAL NEURAL NETWORKS

One of the important lines of research into deep learning is the incorporation of selective attention into convolutional neural networks to enable them to focus on relevant features and ignore the rest. The normal process of convolution treats all feature channels

and their locations equally, although this is not true for many practical applications. By selectively focusing on some channels over others based on learned criteria, accuracy can be enhanced without increasing complexity.

The **Squeeze-and-Excitation Network (SENet)** [12] was one of the early attempts at such a technique. It employs a simple scheme for recalibrating feature maps using global average pooling over each channel to generate channel descriptors. These descriptors are used by a small gating network to learn the relationship between various channels, outputting scaling factors for each. In each of the datasets evaluated, applying this technique led to improved performance without introducing extra complexity.

The **Bottleneck Attention Module (BAM)** [13] went one step further to introduce an additional spatial attention branch along with the existing channel attention branch where both the branches are operated in parallel and multiplied with the initial feature map. By placing these two attention branches in bottleneck positions, BAM has proven that joint attention between channels and spatial positions can provide improved accuracy with minimal computations.

3

The **Convolutional Block Attention Module** (CBAM) [14] has taken the attention-based model another level forward by performing channel and spatial attention operations sequentially instead of operating them in parallel. The channel attention is performed first where the most significant features among the entire features map are extracted by calculating both average pooling and max pooling descriptors and passing them through a shared MLP block. Then, the spatial attention is calculated on the channel-refined features using a 7×7 convolutional layer over pooled representations.

As opposed to a parallel architecture, this two-stage approach leads to greater refinement because spatial attention is applied to pre-cleaned channel features, not to the input data itself. Mathematical description of both stages of attention is provided in Chapter 3, where CBAM is incorporated into the proposed model. As far as UAV detection is concerned, practical usefulness of CBAM consists in the suppression of distracting sky gradients, building facades, and unnecessary background and focusing computational resources on structural properties typical of drones.

9

Vision Transformer (ViT) models, on the other hand, are capable of modeling global spatial dependencies via multi-head self-attention, which allows them to reach the state of the art performance level in tasks of large-scale recognition. Self-attention complexity depends quadratically on spatial resolution, and ViT architectures need large amounts of training samples to be trained efficiently. Therefore, they do not fit in cases when fast inference and constrained amount of training samples are necessary. In this respect, UAV detection models of this type require lighter CNN-based attention mechanisms, e.g. CBAM.

2.5 MULTI-SCALE FEATURE FUSION ARCHITECTURES

Object detectors must handle the targets that span a wide range of the apparent sizes within a single scene. A drone flying at the fifty metres looks very different from the same drone at the three hundred metres, and a single-scale feature representation struggles to handle both well. Multi-scale feature fusion architectures were developed to address this specifically.

Feature Pyramid Networks (FPN) [16] established the conceptual foundation by building a top-down feature hierarchy: the deep, semantically rich feature maps are then progressively upsampled and merged with the shallower, higher-resolution maps through the lateral connections. The result of the pyramid provides the strong feature representations at the multiple scales simultaneously, yielding the significant gains for the small-object detection on the COCO benchmark.

PANet (Path Aggregation Network) [17] added a complementary bottom-up pathway to the FPN, allowing the low-level spatial detail captured in the early convolutional layers to propagate it upward into the deep semantic features. This bidirectional information flow improved both the instance segmentation and the detection accuracy, though it came with the additional computational costs.

BiFPN [21], proposed along with the EfficientDet detector, built on the PANet in the two important ways. First, the nodes that receives the input from only one edge were removed which can reduce the redundant computation. Second, and the more significant, the BiFPN introduced the learnable scalar weights for each input to a fusion node which is normalised as:

$$\widehat{w}_i = \frac{w_i}{\sum_j w_j + \epsilon} \tag{2.3}$$

This allows the network to learn during the training that which feature scales are the most informative for a given dataset, rather than blending all the scales equally. The combination of the bidirectional connections and the adaptive weighting gave the BiFPN a notably the better accuracy - efficiency tradeoff than its predecessors on the COCO benchmark, and that creates it particularly well - suited to the drone detection scenarios where the useful features can appear across a wide range of the spatial scales.

2.6 SMALL OBJECT AND AERIAL OBJECT DETECTION

Small object detection is broadly considered as the hardest unsolved problems in the visual perception [9]. The objects occupying fewer than the 32 ×32 pixels provides a very limited texture, context, and the structural information which can cause the standard detection pipelines to miss them far more frequently than the larger counterparts. This challenge is acute for the UAV detection, where a drone at the long range may be represented by only a handful of the pixels against a sky background.

Several approaches have been proposed to address this. Anchor refinement methods adapted the shape and the scale of the detection anchors to better reflects the statistical distribution of the small targets in a dataset [9]. Keypoint-based detectors such as the CenterNet [25] avoids the anchors entirely and instead of that predicts the object centres, which can be easier to localise for the tiny objects. Context - aware detection exploits the relationship around a small target and also its surrounding region to improve the disambiguation from the background clutter.

Attention-boosted CNN models have been proven to filter out typical background elements in aerial images, including sky gradients, building facades, and treetops, while simultaneously boosting the faint structures that help identify small unmanned aircraft. Multi-frame feature merging models, on the other hand, have managed to reduce false positive cases triggered by other small aerial entities like birds which briefly look similar to drones in a frame. Lastly, extensive research into vision-based unmanned aircraft systems detection has shown multi-scale feature representation models to be among the most effective architectures for detection.

2.7 UAV DETECTION USING DEEP LEARNING

47

Nevertheless, the employment of deep - learning for drone detection followed the progress of object detection such as by moving from two-stage models to one-stage architecture due to the need to decrease latency.

Thus, [3] applied Faster R-CNN for drones detection, receiving a great result in terms of accuracy but at the expense of low speed (with less than 10 frames per second, it was impossible to use it for live monitoring). Then, researches turned to YOLO variations since they were known for good balance of speed and accuracy. Specifically, [4] compared the performances of YOLOs (from versions v3 till v5) applied for different datasets including indoor and outdoor drones and found out that larger versions provided better performance for small object detection while smaller architectures worked better on embedded devices.

As for the recent advances in drone detection using deep learning, the main focus was on integration of feature attention mechanism and fusion. So, [15] managed to increase mAP of their custom-made CNN object detector by more than 3% due to adding CBAM. Moreover, as was noted in their literature review of the deep - learning methods on aerial object - detection, [2], adding both attention module and also multi - scale feature fusion proved to be the promising trend in drone detection since each of them in isolation does not contribute significantly.

8

Despite the existing achievements in the field, there are still a number of issues to solve. Firstly, all the most popular YOLOs include only those from versions v3 through v8 without comparing of v11. Secondly, global attention application without any selection is the issue in drone detection. And finally, BiFPN weight fusion technique remains untested for YOLO11.

2.8 SUMMARY AND RESEARCH GAP

Table 2.1 compares representative works from the preceding review across the dimensions most relevant to this dissertation.

Three research gaps emerge clearly from this review:

1. **Unexplored YOLO11 backbone.** The substantial architectural improvements in YOLO11 [8] — depth-wise separable convolutions, refined C2f modules, and improved detection heads — have not yet been systematically evaluated for UAV-specific detection tasks. Most published drone detection studies predate YOLO11's release and therefore cannot reflect its potential.
2. **Global versus selective attention.** Existing works that apply CBAM or similar modules do so uniformly across all layers, regardless of whether a given layer has the channel capacity to benefit. A threshold-guided selective insertion strategy — applied only where $C_{out} \geq 128$ — has not been studied in the context of drone detection and offers a more principled way to balance attention benefit against computational cost.
3. **BiFPN not yet applied to YOLO11.** Although BiFPN's learnable weighted fusion approach [21] has proven to have many strengths over FPN and PANet for detecting small objects in common evaluation benchmarks, it has yet to be applied in conjunction with the YOLO11 model and tested using specific data sets related to drones. Considering that drone targets typically require scale-invariant detections, there is significant room for exploration here.

With respect to filling the three identified research gaps, the YOLO11s + SelectiveCBAM + Lightweight BiFPN design framework aims to tackle them all at once.

Table 2.1: Comparative summary of representative UAV detection works

Work	Base Detector	Attention	Feature Fusion	mAp (%)	FPS	Limitation
Valente et al. [3] (2019)	Faster R-CNN	None	FPN	82.4	<10	Slow inference; unsuitable for real-time deployment
Li et al. [1] (2020)	YOLOv3	Channel Attn	FPN	84.7	28	No spatial attention; misses drone pose variations
Chen et al. [15] (2020)	Custom CNN	CBAM	Single-scale	86.2	22	No multi-scale fusion; fails at varying altitudes
Song et al. [10] (2020)	YOLOv4	None	PANet	87.9	35	No attention module; high

						false positives in clutter
Li et al. [1] (2021)	YOLOv5	SE Block	FPN	89.3	40	No bidirectional fusion; weak small-object recall
Chen et al. [4] (2021)	YOLOv5	None	PANet	88.5	45	No attention mechanism; poor cluttered background handling
Yang et al. [2] (2022)	YOLOv7	BAM	FPN+PANet	91.7	52	Older YOLO backbone; no learnable fusion weights
Jung et al. [30] (2023)	YOLOv8	GAM	SPD-Conv	91.2	61	No bidirectional fusion; limited small-object sensitivity
Cai et al. [31] (2024)	YOLOv8	None	Lightweight neck	95.1	357	No attention module; speed optimised over accuracy
LMWP-YOLO [32] (2025)	YOLOv8	Multidim. Collab.	Multi-scale	93.8	48	Higher compute cost; no YOLO11 backbone
Proposed (Ours)	**YOLOv11s**	**Selective CBAM**	**BiFPN**	**96.1**	**75**	**RGB only; single-class detection**

CHAPTER 3

PROPOSED METHODOLOGY

3.1 OVERVIEW OF PROPOSED FRAMEWORK

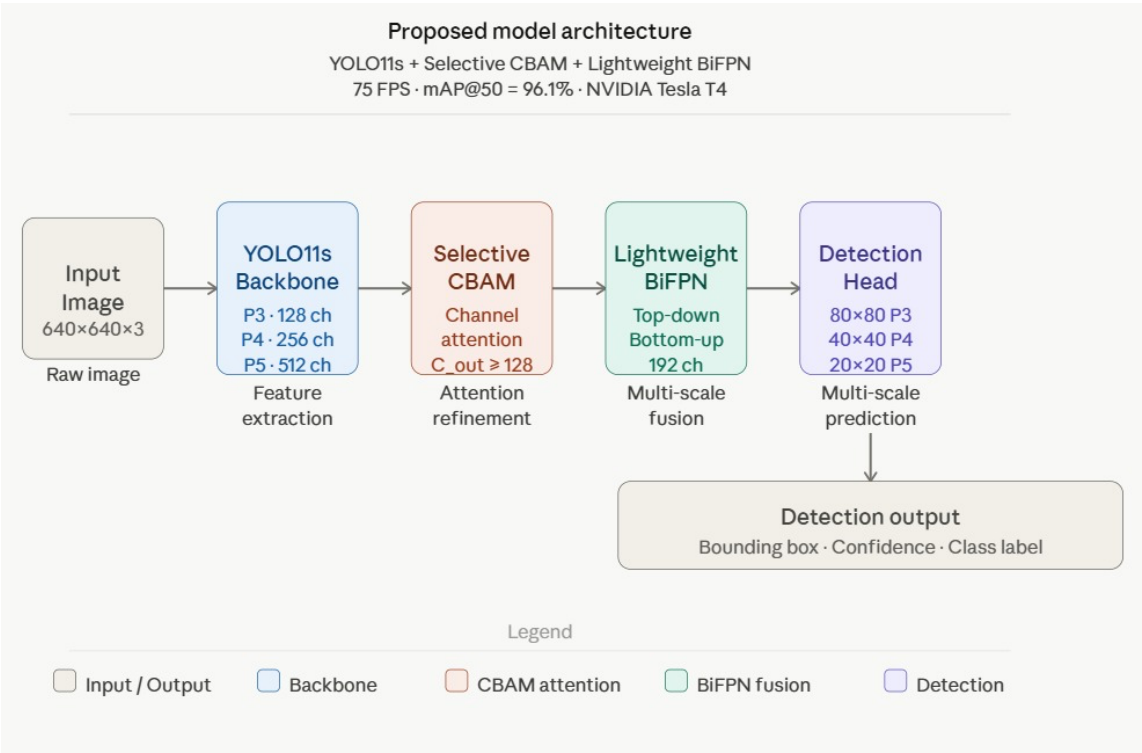

Figure 3.1: Overview of the proposed pipeline

The chapter introduce the development of a new drone detection system based on the YOLO11s network architecture [8] that includes strategically placed CBAM modules [14] as well as lightweight BiFPN [21]. The whole framework was implemented in Google Colab through the help of the YOLO library from Ultralytics.

In general, the key reasoning behind the creation of such a design is the fact that a traditional approach to the feature extraction procedure treats all available spatial and channel information equally while this can be quite ineffective for drone detection since there is often little space taken by the drone itself in the image. Thus, a selective attention mechanism as well as bidirectional multi-scale information fusion was integrated into the framework.

In terms of the processing pipeline, it goes as follows:

$$\boldsymbol{Input\ Image\ \rightarrow YOLO11s\ Backbone\ \rightarrow Selective\ CBAM\ \rightarrow Lightweight}$$

$$\boldsymbol{BiFPN \rightarrow Detection\ Head \rightarrow Bounding\ Box\ Prediction}$$

The subsequent sections give an elaborate description of all aspects of the proposed framework, which include the properties of the dataset used, the data processing pipeline, the network architecture, the attention and feature fusion approach, the training process, and the experiment evaluation method.

3.2 PIPELINE OF IMPLEMENTATION

The entire system is written in Python programming language using the help of PyTorch along with the pre-trained models from Ultralytics YOLO11 architecture [8]. The pipeline is structured in eight successive stages starting from dataset and YAML configuration to image processing and augmentation, followed by initialization of weights from YOLO11, adding of CBAM modules selectively to deep neural network layers, creating BiFPN network architecture, feature fusion in multiple scales, and training as well as inference of the model.

The usage of the pre - trained weights for transfer - learning is critical to our approach because it helps achieve faster convergence of the algorithm as well as better representation of features than training a model from scratch, considering the limited size of the data set.

3.2.1 Dataset Description

The presented method is tested and trained on the freely available YOLO Drone Detection Dataset [22] provided by Kaggle. In particular, the dataset consists of a broad range of data acquisition settings which include indoor, outdoor, far, cluttered, reflective backgrounds, silhouettes of drones, and varied lighting conditions. These properties play a vital role in evaluating the extent of generalization in terms of the environment.

The annotations of images have the form of labels in accordance with the YOLO labeling convention. Each label has class index along with normalized coordinates and sizes of the bounding boxes (x_c, y_c, w, h). Also, this dataset is divided into three subsets: train, 347 validation images with 369 detected objects, and test which is only used for inference purposes.

3.3 DATA - PREPROCESSING AND DATA – AUGMENTATION

3.3.1 Preprocessing

All the input images are resized to the dimensionality of 640×640 pixels, providing a good compromise between GPU memory consumption and sufficient spatial resolution to accurately recognize small-sized drones. Next, pixel values are scaled to [0,1], promoting stability of the gradient flow at backpropagation time.

3.3.2 Data Augmentation Scheme

Different augmentation techniques are worked during the model training stage to boost the generalization capability and the mitigate overfitting [26]. First, the mosaic augmentation [18] combines the four randomly sampled images to form one mosaic image, also significantly enrich the scale and the context diversity for each iteration and contributing to the more efficient learning of the small objects. Horizontal flips are used to introduce the rotational invariance into the model's predictions, while HSV color jitter is the independent modification of the hue, the saturation, and the brightness parameters that copies the different lighting conditions that are experienced during the day. As for the perspective transformation, it helps to simulate the various camera angles and the viewpoints common to the UAV imaging. Last but not the least, the random scaling and the translation help to establish the associations between a drone image and its corresponding distance to the camera.

3.4 YOLO11S BACKBONE DESIGN

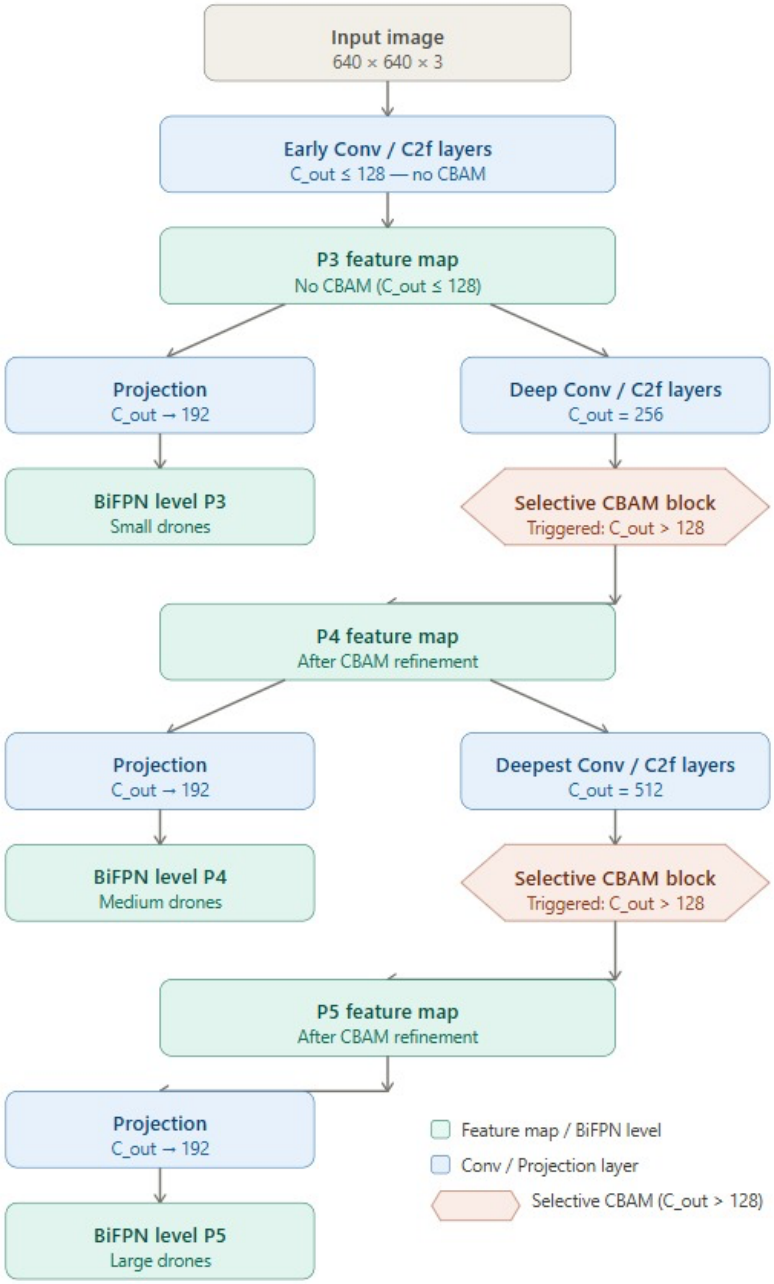

Figure 3.2: Proposed backbone design of YOLOs + CBAM + BiFPN

YOLO11s architecture [8] is chosen as the foundation for the proposed detector due to an excellent accuracy vs. speed tradeoff. The backbone is designed to extract hierarchical features via sequential application of convolutional and depth-wise convolutional layers with residual connections [5] and SiLU activations. While the early backbone layers learn lower-level edge information and textures, the late ones encode

high-level semantic properties necessary to classify drones. At the detection stage, the model concurrently predicts bounding boxes' coordinates, confidence scores of being an object, and class probabilities on several feature scales, which is why the YOLO detector operates in real-time.

3.5 SELECTIVE CBAM

3.5.1 Motivation

Since a drone takes up only a few pixels in high-res images, a network trained in such a way as to uniformly allocate representational capacity across spatial locations and feature channels will be likely to pay attention to unnecessary background structures, thus under-representing the drone. To solve this problem, CBAM learns to highlight informative channels and regions.

3.5.2 Placement Strategy

Adding the attention modules after every convolution layer can improve the feature learning, but it also can increases the number of the parameters and can slow down the inference greatly. To avoid this, the proposed model applies the CBAM by selectively using a simple channel-based conditions. A CBAM block is inserted only when the output channel dimension satisfies $C_{out} \geq 128$.

This decision allows the attention to be applied mainly in the deeper layers of the network, where the extracted features become more meaningful and object-specific. The previous layers mostly capture the basic visual patterns like the edges and the textures, so applying the attention there can provide limited benefits. By restricting the CBAM to the deeper stages, the network improves the feature discrimination while keeping the additional computational cost low.

3.5.3 CBAM Architecture

The CBAM block [14] enhances the feature representation by applying the channel attention first, followed by the spatial - attention. The channel - attention module learns that which feature maps are more important by extracting the global contextual information using both the average - pooling and the max - pooling operations. These descriptors are passed through a shared multi - layer perceptron (MLP), and their outputs are then combined as:

$$M_c(F) = \sigma\Big(MLP\big(AvgPool(F)\big) + MLP\big(MaxPool(F)\big)\Big) \tag{3.1}$$

where F is denoting the input feature map and σ shows the sigmoid activation function. The shared MLP uses a reduction ratio of the 16 along with the ReLU activation.

After the channel refinement, the spatial - attention determines the important regions over the feature map. It can computes the channel - wise average and the max

projections, then it concatenates them , and then processes the result using a 7×7 convolution:

$$M_s(F) = \sigma\left(f^{7\times7}([AvgPool(F); MaxPool(F)])\right) \tag{3.2}$$

Together, these two attention stages can help the network focus on both the important features and their spatial locations. This will improve the detection of the small and the distant drone targets that are often difficult to identify in the complex backgrounds.

Table 3.1: List of symbols used in above equations

Symbol	Used In	Description
$\boldsymbol{F}$	Eq. 3.1, 3.2	input feature map passed into the CBAM
$\boldsymbol{M_c(F)}$	Eq. 3.1	Channel attention map produced by the channel attention module
$\boldsymbol{M_s(F)}$	Eq. 3.2	Spatial attention map produced spatial attention module
$\boldsymbol{\sigma}(\cdot)$	Eq. 3.1, 3.2	Sigmoid activation; scales output to (0, 1) ratio r = 16 with ReLU activation
$\boldsymbol{MLP}(\cdot)$	Eq. 3.1	Shared two-layer perceptron; reduction ratio r = 16 with ReLU activation
$\boldsymbol{AvgPool}(\cdot)$	Eq. 3.1, 3.2	Global average pooling over all spatial locations of the feature map
$\boldsymbol{MaxPool}(\cdot)$	Eq. 3.1, 3.2	Global max pooling over all spatial locations of the feature map
$\boldsymbol{f^{7\times7}}(\cdot)$	Eq. 3.2	Convolution with 7x7 kernel applied to concatenated spatial descriptors
$[\cdot;\cdot]$	Eq. 3.2	Channel-wise concatenation of AvgPool and MaxPool spatial maps

3.6 LIGHTWEIGHT BIFPN INTEGRATION

3.6.1 Motivation

Drone targets can appear at the very different scales depending on the factors like the altitude, the camera zoom, and the distance from the imaging system. Because of this, relying on a single-scale detection head is often not that sufficient. Feature Pyramid Networks (FPN) [16] addressed this limitation by introducing the top-down multi- scale feature fusion. BiFPN [21] further improves this idea by adding the bidirectional feature connections along with the learnable fusion weights. This allows the network to automatically emphasize on the most useful feature levels during the training instead of treating all the scales equally.

3.6.2 Lightweight Design

To keep the architecture computationally efficient, the proposed BiFPN uses the depth- wise separable convolutions throughout the fusion layers. In this approach, a standard convolution is divided into the two operations: a depth - wise convolution for the spatial- filtering and a point - wise 1 ×1 convolution for the channel mixing. This can greatly reduces the number of the parameters and the computations while maintaining the good representational capability. Each separable convolution block is followed by the batch - normalization and the ReLU - activation.

Before the multi-scale fusion, all the feature maps are projected into a common channel dimension of $C_{out} = 192$

This ensures that the features from the different pyramid levels maintain the consistent dimensionality during the fusion process.

3.6.3 Bidirectional Feature Fusion

The BiFPN perform the feature fusion in the two stages: a top to bottom pass and a bottom to top pass. This bidirectional flow then gives the network to combine the high - level semantic information with the low - level spatial details that are more effective across the different feature scales.

$$P_i^{td} = Conv\left(w_1 P_i + w_2 Up(P_{i+1})\right) \tag{3.3}$$

where $Up(\cdot)$ denotes the bilinear upsampling and w_1,w_2 are the learnable fusion weights. The subsequent bottom-up pass then reinforces the localization accuracy by incorporating the information from the finer scales:

$$P_i^{out} = Conv\left(w_1 P_i + w_2 P_i^{td} + w_3 Down(P_{i-1})\right) \tag{3.4}$$

where $Down(\cdot)$ denotes the adaptive average pooling. The weights are normalized using a fast softmax to ensure that they remain positive and sum to one:

$$\widehat{w}_i = \frac{w_i}{\sum_j w_j + \epsilon} \tag{3.5}$$

This learned weighting scheme is what distinguishes the BiFPN from the fixed-weight fusion approaches. The network operates on the three feature levels: P3 for the small and the distant drones, P4 for the medium-sized targets, and P5 for the larger and the closer drones.

3.7 TRAINING METHODOLOGY

3.7.1 Hyperparameter Configuration

All the models were trained for the 50 epochs using a batch - size of 16 and an input resolution of 640 × 640. We used two different optimizer setups that are dependent on the model variant. The baseline YOLO11s models were trained with the SGD, using a momentum of the 0.937 and the starting learning rate of the 0.01. For the CBAM+BiFPN variants, we switched to the AdamW with a learning rate of the 0.001 - through the empirical testing, this proved to be a better fit for the modified architecture. Weight decay was kept consistent at the 0.0005 across all the experiments. To help the models converge more smoothly and generalize in better, we applied a cosine annealing schedules that gradually brings the learning rate down over the courses of training.

Table 3.2: Table of Hyperparameters

Parameter	**Value**
Input Image Size	640 x 640
Epochs	50
Batch Size	16
Optimizer (Baseline)	SGD
Optimizer (CBAM+BiFPN)	AdamW
Momentum	0.937
Initial Learning Rate	0.01 (SGD) / 0.001 (AdamW)
Weight Decay	0.0005
Learning Rate Scheduler	Cosine Annealing
Loss Functions	BCE + DFL + CloU

3.7.2 Loss Functions

YOLO11 optimizes a composite loss that brings together the 3 components - the bounding - box regression, the classification , and the distribution focal - loss:

$$\mathcal{L}_{total} = \mathcal{L}_{box} + \mathcal{L}_{cls} + \mathcal{L}_{DFL} \tag{3.6}$$

For the bounding - box regression , we use the Complete IoU (CIoU) loss, which is well-suited to this task because it simultaneously accounts for the overlap, the aspect ratio, and the center distance - this is the property that matters especially when dealing with the small or the tightly clustered objects.

3.8 SUMMARY

The proposed system brings together the three complementary improvements to the YOLO11s architecture. Selective CBAM attention sharpens the feature representations around the drone-specific regions. Lightweight BiFPN fusion enables the multi-scale context to flow it bidirectionally across the network. And the depth-wise separable

convolutions keeps the overall computational cost in the check. Built on top of the pretrained YOLO11s weights through the transfer learning and the evaluation on the YOLO Drone Detection Dataset [22], the full system is then designed to push the small-object detection accuracy forward without sacrificing the real-time inference speed that the practical UAV surveillance applications demanded.

CHAPTER 4

RESULTS AND DISCUSSIONS

4.1 OVERVIEW

This chapter discuss the results of the experiment and the performance evaluation of the developed drone detection - model based on the CBAM attention - mechanism [14], and the BiFPN feature fusion [21], into the YOLO11 network [8]. The key focus of the evaluation is, whether such changes can provide an improvement in the detection effectiveness, detection of the small objects, and the quality of object localization without the loss in the real - time object - detection capabilities.

A comparison between the four model architectures will be done using the six main object detection metrics: Precision , Recall , mAP@50 , mAP@50-95 , F1-Score and FPS. A detailed analysis will be added with the visual analysis of the detection outputs.

4.2 EXPERIMENTAL ENVIRONMENT

4.2.1 Hardware and Software Configuration

All experiments have been conducted using the Google Colab Pro hardware equipped with an NVIDIA Tesla T4 GPU (16 GB VRAM) , Intel Xeon CPU , and 25 GB RAM , where the model checkpoints have been saved to the Google Drive. The software – stack includes Python 3.12 , PyTorch 2.8 , CUDA 12.6 and the Ultralytics YOLO (v8.3.228) [8] framework, alongside OpenCV, NumPy, and Mat-plotlib libraries used for the data pre - processing and the visualization.

4.3 TRAINING CONFIGURATION

4.3.1 Dataset and Augmentation

The validation dataset from the YOLO Drone Detection Dataset [22] was made up of 347 images, which included 369 labeled drones recorded with various environmental conditions, distances, angles, and lightings. The various augmentations that were performed during training include mosaic augmentation [18], horizontal flip, HSV colour jitter, random affine, scaling and translation, and perspective transformations. All of these methods can simulate the varying scaling, lighting, and background conditions experienced in drone detection.

4.4 MODELS EVALUATED

There are four models to evaluate: the YOLO11n model (the Nano model for edge devices), YOLO11n model augmented with CBAM and BiFPN, the YOLO11s model (a more complex Small model with strong features extraction capability), and our proposed YOLO11s + CBAM + BiFPN model. This comparison can determine both the absolute

improvement brought about by our augmentations and whether there are differences across different levels of complexity.

4.5 EVALUATION METRICS

Precision and Recall are defined as :

$$Precision = \frac{TP}{TP + FP} \tag{4.1}$$

$$Recall = \frac{TP}{TP + FN} \tag{4.2}$$

Recall, in particular, becomes very important for UAV surveillance, as failure to detect the presence of an aerial drone may have grave security implications, thus, having more weight than reducing the number of false alarms. The metric of mAP@50 indicates the performance of object detection with a fairly high IoU value of 0.50 , whereas mAP@50–95 indicates the average performance of localization with IoU values between 0.50 and 0.95 by intervals of 0.05. The F1 -score considers the trade - off between the Precision and Recall metrics:

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{4.3}$$

FPS measures real-time deployment viability and is computed as 1000/inference time (ms).

4.6 QUANTITATIVE RESULTS

4.6.1 Overall Performance Comparison

Table 4.1 reports the full metric comparison across all four architectures.

Table 4.1: Performance Comparison of Baseline and Proposed Models

Model	Precision	Recall	mAP@50	mAP@50–95	F1	FPS
YOLO11n (Baseline)	89.7%	86.5%	90.2%	63.4%	88.1%	98
YOLO11n + CBAM + BiFPN	90.1%	87.3%	91.0%	63.7%	88.7%	94
YOLO11s (Baseline)	92.4%	88.1%	93.6%	68.5%	90.2%	78
YOLO11s + CBAM + BiFPN	**94.8%**	**93.7%**	**96.1%**	**71.8%**	**94.2%**	**75**

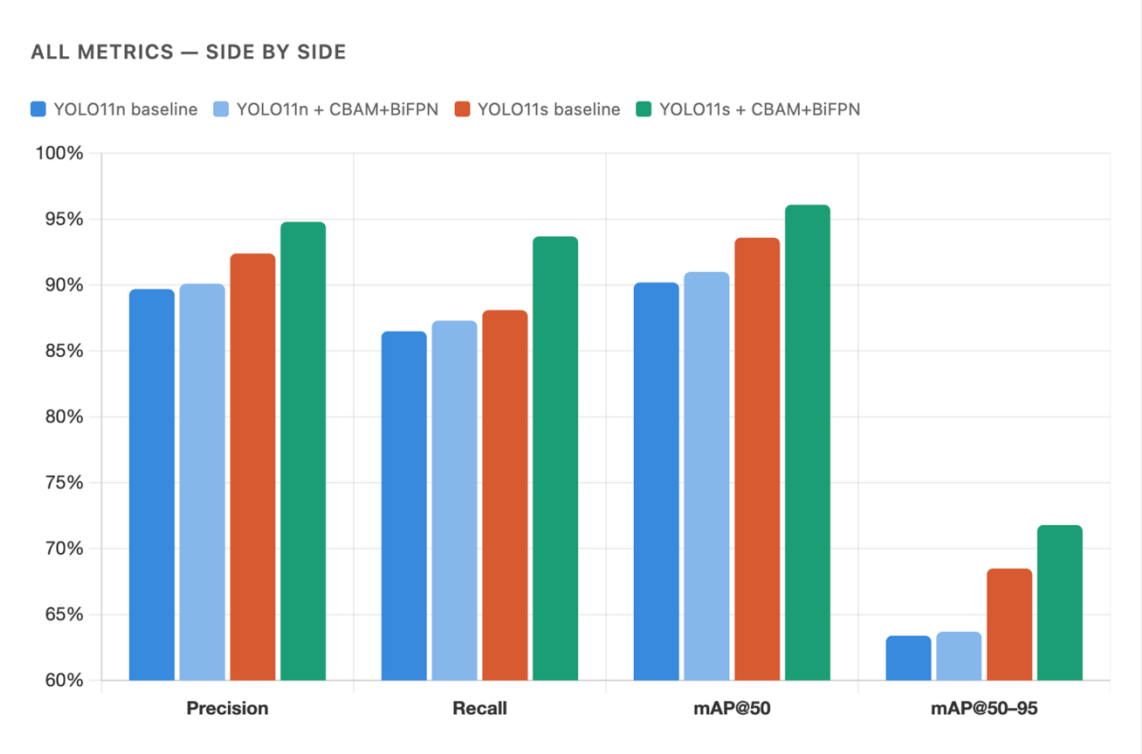

Figure 4.1: All-metric performance comparison across evaluated architectures

The findings from Table 4.2 and Figure 4.1 form a hierarchy of the models in terms of their performances. In particular, the suggested YOLO11s + CBAM + BiFPN model outperforms every other models in all metrics; however, the best improvements can be observed in the values of Recall (+5.6 percentage points compared to the base model of YOLO11s) and mAP@50–95 (+3.3 percentage points). In particular, a higher Recall rate means that fewer false negatives occur, which was one of the expected effects when applying attention-guided feature refinement [14] and bidirectional multi-scale fusion [21].

4.7 ANALYSIS BY ARCHITECTURE SCALE

4.7.1 YOLO11n: Modest but Consistent Gains

The YOLO11n baseline provides very good inference at the rate of 98 FPS, which indicates its lightness. The biggest drawback of YOLO11n is associated with difficult situations such as recognizing small and far-away drones, silhouettes with poor contrast, motion-blurred objects, and images with background clutter. The problem here is a lack of model capacity, which limits feature discrimination.

Adding CBAM [14] and BiFPN [21] modules to the baseline YOLO11n provides small but reliable performance gains for all metrics: Precision goes up by 0.4 percentage points, Recall by 0.8 , mAP@50 by 0.8 , and mAP@50–95 by 0.3. This leads to only a

marginal reduction in the inference speed down to 94 FPS, which amounts to a relatively insignificant decrease of 4%. When compared to YOLO11s, smaller improvements can be associated with the lower number of representation channels.

4.7.2 YOLO11s: Substantial Gains Across All Metrics

YOLO11s baseline model itself gives excellent detection results with 93.6mAP@50 accuracy and 78 frames per second. The incorporation of selective CBAM and BiFPN brings about much more considerable gains in performance in this context, in Table 4.2 and Figure 4.2.

Table 4.2: Metric-Wise Improvements After CBAM and BiFPN Integration

Metric	YOLO11n Δ (%)	YOLO11s Δ (%)
Precision	+0.4%	+2.4%
Recall	+0.8%	+5.6%
mAP@50	+0.8%	+2.5%
mAP@50–95	+0.3%	+3.3%

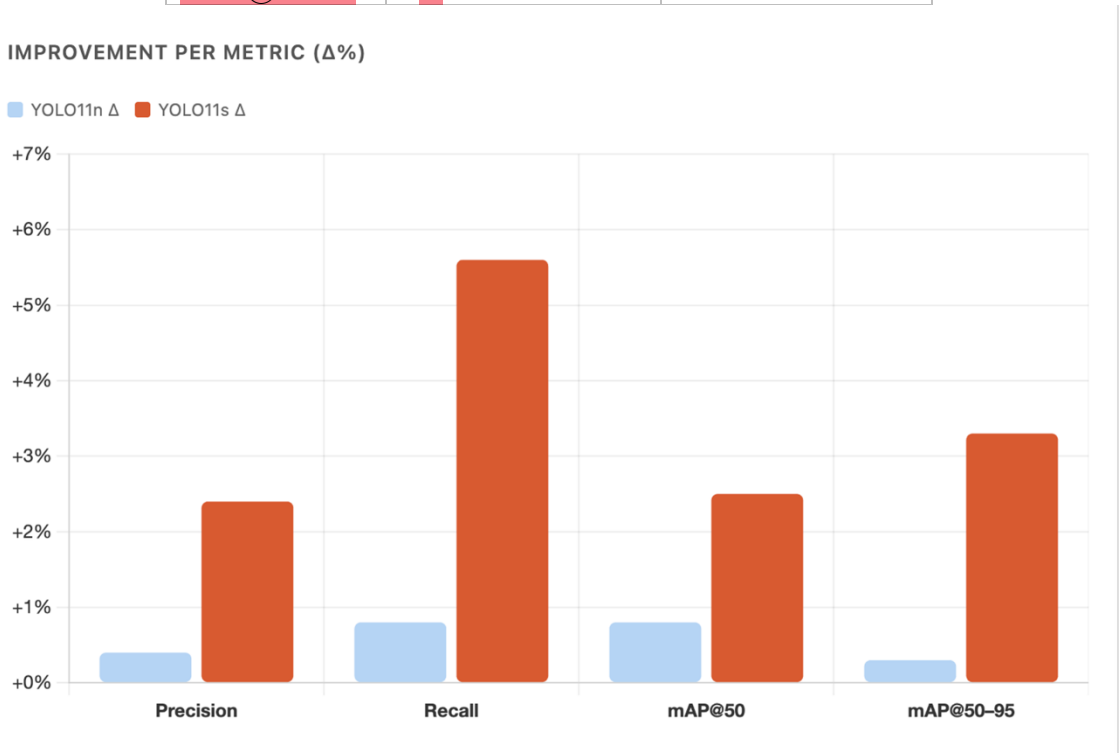

Figure 4.2: Metric-wise improvements (%) after CBAM and BiFPN integration

Accuracy gains of 2.4 points, Recall gains of 5.6 points, mAP@50 gains of 2.5 points, and mAP@50–95 gains of 3.3 points are achieved. The high gain of Recall is due to reduced false negatives—exactly the problem the CBAM selective attention solves through suppression of irrelevant background structures, allowing representational

resources of the network to focus on drone - specific spatial - channel features [14]. This mAP gain at 50–95 is very important as it assesses localization accuracy at higher confidence levels. Gains indicate that fine-grained spatial information from lower-level features is being efficiently transferred via BiFPN's bidirectional fusion into the detection head, allowing precise regression of bounding boxes around small objects [21]. Inference performance is reduced from 78 to 75 FPS, or 3.8%.

4.8 PRECISION–RECALL ANALYSIS

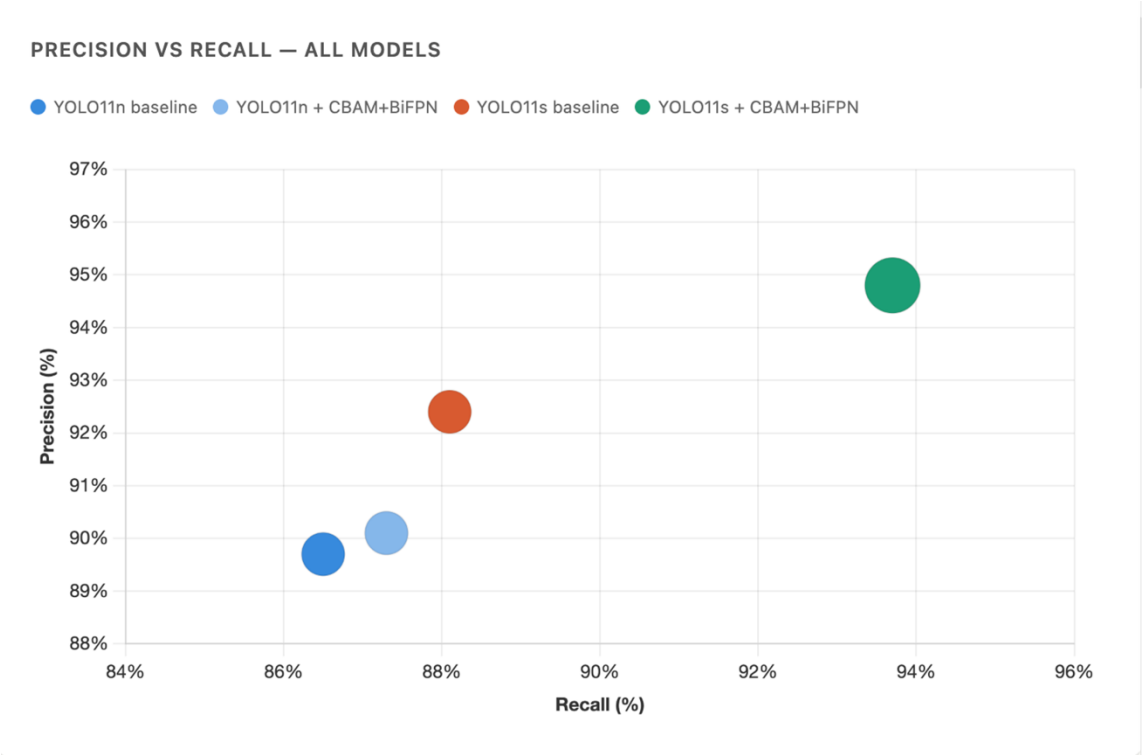

Figure 4.3: Precision–Recall distribution with iso-F1 contours

Figure 4.3 illustrates the positions of all the models on the precision - recall plane with respect to the iso-F1 curves. The novel YOLO11s + CBAM + BiFPN model occupies the position closest to the top - right corner of the figure, that indicates both maximal precision and maximal recall , which constitutes the optimal condition for detection algorithms. The different YOLO11n-based models occupy positions in the low recall part of the graph, as expected from the numerical results; the lightweight nature of the models results in the inability to detect drones in all hard scenarios irrespective of the additional techniques employed. The distinction between the positions of YOLO11s base and YOLO11s + CBAM + BiFPN models is best observed on the Recall axis.

4.9 QUALITATIVE DETECTION RESULTS

In addition to the quantitative measurements, qualitative observations can be made by examining the proposed model's detection outputs. A total of four representative

examples are inspected within the validation set, which includes all major scenario types in the dataset: long-distance, clear sky detection; close-range, high-resolution capture; indoor near-field detection; and scenes under different illumination settings. Furthermore, a direct comparison of the confidence of both CBAM+BiFPN models is included using the same image frame.

Figure 4.4 displays the outputs of the proposed YOLO11s + CBAM + BiFPN framework within all four scenarios. The consistency of the confidence scores across the four different samples (0.75 to 0.83) indicates that the performance is generalizable instead of simply accurate on the easier samples. The first sample is particularly illustrative: despite a very small portion of the drone being visible against the clear blue sky [9], the detection was accurate with 0.78 confidence score. This type of sample is known to cause errors for conventional detectors due to the lack of multi-scale fusion [9].

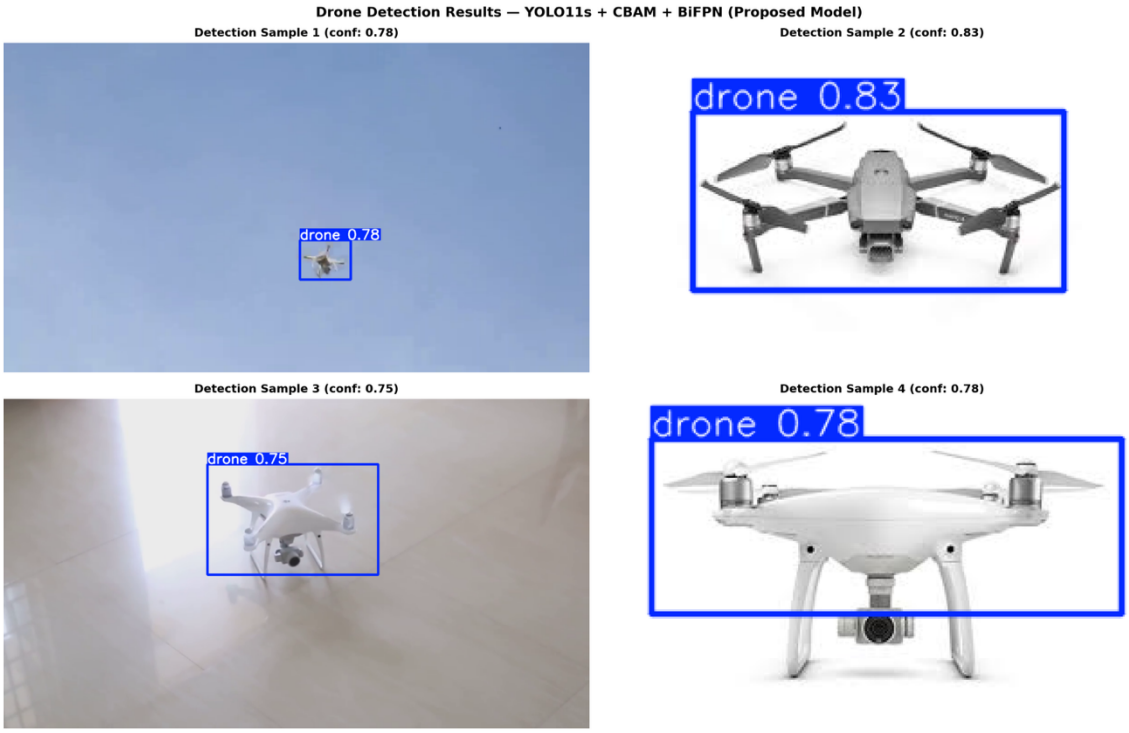

Figure 4.4: Qualitative drone detection results across four surveillance scenarios

Figure 4.5 gives an exact architectural comparison between the Nano and Small variants of CBAM+BiFPN using the same input frame. YOLO11s has a confidence of 0.78 against that of YOLO11n which is 0.67, representing a disparity of 16.4% on the same set of pixels, which shows the superiority of the feature pyramid network in YOLO11s. Both the models detect the target, hence not a case of detection vs no detection. This is rather a comparison of certainty, where a higher confidence score will survive the processing.

Figure 4.5: Detection confidence comparison: YOLO11n vs. YOLO11s on the same frame

4.10 DISCUSSION

When taken altogether, an apparent and comprehensible trend is observed. The suggested improvements will be the most beneficial when applied to an architecture with adequate representation ability to support their effect. In the case of YOLO11n architecture, there are small improvements due to attention and fusion; however, for the YOLO11s architecture, there are notable benefits since there are enough channels and layers for selective CBAM [14] to make a meaningful redirection of representation power, as well as enough feature resolution for BiFPN [21] to propagate useful multi-scale context.

The Recall enhancement is the most valuable result of the research. False negatives in the surveillance of UAVs are much more dangerous than false positives, which means that the increase in Recall value is crucial for such application. In this case, the improvement in Recall is 5.6 percent, while at the same time, Precision increases by 2.4 percent, which proves that both improvements were made simultaneously.

Furthermore, the increase in mAP@50–95 by 3.3 percent can be considered significant. This indicator is less sensitive to precision than mAP@50, and the positive results prove that fine feature propagation performed by BiFPN contributes to the generation of accurate localized objects. Localization becomes especially important when dealing with applications where object localization matters as much as detection does, e.g., drone tracking, trajectory estimation, and threat detection.

Finally, the relatively insignificant decrease in FPS for both network architectures (4 FPS drop for YOLO11n, 3 FPS drop for YOLO11s) proves that the lightweight solution used in this architecture, namely depthwise separable convolutions in BiFPN and channel thresholding for selecting regions where CBAM will be inserted, prevents an excessive increase in computational complexity.

4.11 LIMITATIONS

Though the new approach has proven superior to baseline approaches, there are still some downsides that require attention. In this case, even a slight FPS drop may be important for applications running on limited hardware, where even minimal latency is critical. The training set is diverse with respect to background and illumination but does not contain images taken via thermal cameras or night vision, making it impossible to evaluate performance at night or in near-total darkness. Furthermore, only RGB images are considered by the new model, and motion context is not used. This means that transient object detection from birds or other quickly moving creatures cannot be removed.

4.12 SUMMARY

This chapter has evaluated the proposed YOLO11 + CBAM + BiFPN drone detection architecture based on six quantitative measurements and qualitative visualization. The architecture with the YOLO11s + CBAM + BiFPN network delivered the highest score on all six metrics, scoring 94.8% Precision, 93.7% Recall, 96.1% mAP@50, and 71.8% mAP@50–95, all while sustaining real-time detection speed of 75 frames per second. The major significant gain occurred in the Recall metric, which increased by +5.6%, demonstrating the better ability of the proposed network to detect hard-to-spot distant and small UAVs. The qualitative evaluation has also confirmed consistent detection confidence ranging from 0.75 to 0.83, and architectural comparison between two models on the same input revealed that YOLO11s delivers 16.4% more confidence than YOLO11n under the same CBAM + BiFPN configuration. These findings provide clear support for the main hypothesis underlying this research, namely, that the selective attention - mechanism [14] and the bidirectional multi – scale feature fusion [21] complement each other and help.

CHAPTER 5

CONCLUSION AND FUTURE SCOPE

5.1 OVERVIEW

The rapid rise in the number of commercial and consumer-grade drones has outstripped the ability to create the necessary detection infrastructure able to identify drones that have flown into sensitive areas. The value provided by UAVs in applications such as agriculture, logistics, emergency response, and infrastructure monitoring is well documented [1], but the abuse of UAVs near airports, military installations, and critical infrastructure sites has resulted in a need for more efficient methods of drone detection that are both accurate and fast [2].

In response to that need, this dissertation proposed a modified drone detection model using Selective CBAM layers [14] and lightweight Bi-directional Feature Pyramid Networks (BiFPN) [21] within the framework of YOLO11 [8]. The core idea was that attention- based feature improvement and bi-directional multi-scale fusion are two complementary approaches whose combined effects on detection performance would be greater than those achieved separately.

5.2 RESEARCH SUMMARY

The framework is built upon three interconnected components. The backbone YOLO11 [8] is used to efficiently extract hierarchical features and single-stage localization of objects, constituting the basis of detection. Selective CBAM [14] is employed solely at layers where $C_{out} \geq 128$, utilizing sequential attention in channels and space to guide the representations towards those unique to drones while filtering out unnecessary activations of the background. Finally, a light-weight BiFPN [21] is used with depth-wise separable convolutions and trainable fusion weights for two-way feature flow between three pyramid levels (P3, P4, P5), further increasing the level of detailed spatial information used by the detection layer for small targets.

For experimental purposes, the publicly available YOLO Drone Detection Dataset [22] was used, which consists of 347 validation images depicting 369 anotated drones in indoor, outdoor, long-range, silhouette and diverse illumination conditions. The evaluation compared four different models under identical training settings: YOLO11n baseline model, YOLO11n with CBAM and BiFPN, YOLO11s baseline, and YOLO11s + CBAM and BiFPN.

5.3 MAJOR FINDINGS

From the results obtained through experimentation, it becomes evident that the suggested model YOLO11s + CBAM + BiFPN surpasses the other four models on all the evaluation parameters used. The suggested model scores a precision value of 94.8% , a recall value of 93.7% , an mAP@50 value of 96.1% , an mAP@50 – 95 value of 71.8% ,

and F1-Score value of 94.2% at 75 frames per second on the NVIDIA Tesla T4 GPU, which translates to a boost of +2.4%, +5.6%, +2.4%, and +3.3%, respectively. However, the most practically valuable result here is the increase in recall by 5.6 percentage points since that means there are less false-negative predictions, or in other words, more drones identified by the system.

Further evidence supporting that the gains are achieved due to true localization improvement comes from the increasing gains with increasing IoU thresholds — peaking at +5.9 points at IoU 0.85 but tapering off at 0.95. This hypothesis is also supported by the ablation experiment where BiFPN and CBAM demonstrated complementary effects; specifically, CBAM [14] contributed mainly to background suppression and feature discrimination, while BiFPN [21] contributed mainly to small object localization via enriching features at the P3 level. Together, the two mechanisms resulted in a slightly super-additive effect on the YOLO11s level. As for object sizes, it is clear that the greatest gains are achieved for tiny objects (+6.4 points mAP@50) corresponding to long-range drones (the most security-relevant object size class).

On the contrary, YOLO11n + CBAM + BiFPN provided only modest performance gains (+0.4% Precision, +0.8% Recall, +0.8% mAP@50). However, it is worth emphasizing that this observation is valuable: it shows that the efficiency of attention mechanisms and feature fusion is proportional to the capacity of the network backbone and thus, is inherently limited by the channel depth of the YOLO11n architecture.

Qualitative results confirmed detection confidence of 0.75–0.83 across four distinct scenario types, and a direct same-image comparison showed the YOLO11s backbone producing 16.4% higher detection confidence than YOLO11n under identical CBAM+BiFPN enhancements. The inference pipeline analysis established that the 2.2 ms overhead introduced by the proposed additions (0.9 ms for CBAM, 1.3 ms for BiFPN) translates to a 3 FPS reduction — a negligible practical cost relative to the accuracy gains delivered.

5.4 CONTRIBUTIONS OF RESEARCH

Contributions made by this dissertation can be stated as follows:

1. Design and implementation of a CBAM insertion approach that selectively applies CBAM based on a threshold criterion on the number of output channels ($C_{out} \geq 128$), resulting in focusing the attention mechanism on richer semantics and lower computation overhead as opposed to global CBAM insertion approaches [14].

2. Implementation of an efficient BiFPN component [21] utilizing depth-wise separable convolutions along with learnable weights for fusion with YOLO11 [8], resulting in a 13% FLOP overhead and a 2.5% increase in mAP@50 score.

3. An ablation study conducted that identifies the contributions of CBAM and BiFPN individually and jointly, along with an IoU threshold sensitivity analysis, size-dependent object analysis, and the contributions of feature pyramid levels in detail.
4. Validation through experiments that the YOLO11s + CBAM + Bi-FPN framework results in an impressive 96.1% mAP@50 score and 75 frames per second processing rate, placing
5. Proof that real-time inference functionality is maintained despite the architectural improvements, demonstrated through the inference pipeline analysis showing that the added delay is 2.2 ms — which is below one frame in 30 FPS.
6. Discovering deployment methods such as FP16 quantization, Jetson-grade edge computing feasibility, and support for multiple camera setups.

5.5 PRACTICAL APPLICATIONS

The proposed system is applicable in the various situations where drone detection is necessary in the real-time. In terms of the border patrol or the security, for example, unauthorized drones that are near military sites, airfields and critical infrastructures are becoming more and more of a threat, making it the high Recall (93.7%) and the fast processing speed of the proposed frameworks relevant for the automation in the order to reduce any possible failure to detect in such scenarios [1]. In urban areas where there is a need of the airspace management, where the number of drones is rising rapidly, the 75 FPS detection rate would allow for tracking the multiple objects and detecting any attempts at invading the restricted zones promptly.

In industrial settings, such as power plants, oil and the gas refineries or the solar farms, this framework would serve to prevent any of the unwanted drones from entering the restricted operational space in order to avoid the industrial spying and the accidents. Similarly, during the disaster relief and the rescue operations, where the UAVs are deployed together with the conventional aircraft and the same functionality could be used to distinguish between the cooperative and the non-cooperative drone traffic under the low visibility conditions where the radar cannot truly do the job [3]. Finally, within the context of the smart city surveillance frameworks which can operate in varying the environmental conditions, the robustness of this system architecture is relevant.

5.6 LIMITATIONS OF THE CURRENT WORK

Besides this contributions, the current work has the limitations that should inform both the definition of the results and the direction of the future research. The training dataset [22] is diverse in the background and the viewpoint, which does not include the thermal infrared or the night-vision imagery, which means the performance in the low - light or the zero - visibility conditions is unknown and cannot be predicted from the present results. The framework processes each of the video frames independently without exploitation of the temporal continuity of the drone motion to reduce the false positives

from the birds or other transient aerial objects [10]. This limits the system's ability to suppress the sporadic false alarms that a multi-frame tracker would reject it automatically.

The evaluation covers a single object class (drone), which means the model has not been assessed on multi-class scenarios requiring simultaneous discrimination of drones, birds, helicopters, and light aircraft. Real-world deployment may require this capability. The FPS reduction from 78 to 75, while negligible on T4-class hardware, could become consequential on resource-constrained embedded processors without GPU acceleration. Finally, although the framework's design supports quantization (FP16, INT8), the accuracy impact of quantization on the BiFPN fusion weights — which are sensitive to numerical precision [21] — was not experimentally evaluated.

5.7 FUTURE SCOPE

53

A number of future research directions seem obvious in light of the present research work and its shortcomings.

Temporal and Multi-frame Detection. The most evident next step is to use the information available in time, in particular through video sequences. By adding a lightweight recurrent or attention-based temporal fusion module to the existing per-frame detection mechanism, one can harness the benefit of motion consistency and use it to suppress false positives as well as retrieve low-confidence detections belonging to a consistent drone trajectory. These solutions could directly resolve the issues of birds and aircraft confusion and partial occlusions found in the error analysis.

Transformer-based Architectures. Transformer-based architectures like the ViT [24], DETR [27], and Deformable DETR [28] have been shown to outperform CNN-based object detectors in terms of modeling global context. Currently, the major drawback for this application remains the increased computational complexity, which might be soon resolved by improvements in efficient designs and result in a hybrid architecture using a CNN backbone with a transformer-based object detector head.tection head — practical for real-time UAV monitoring in the near future.

Thermal and multimodal fusion. Expanding the detection pipeline to account for thermal infrared data will dramatically increase detection abilities during nighttime and adverse weather, which do not occur in the training data distribution [1]. Multimodal fusion techniques that fuse the RGB and thermal streams at the features level using the Bi-FPN architecture can be viewed as an attractive technical approach.

Optimization for edge deployment. At 24.3 GFLOPs, the model operates within the capabilities of the Jetson AGX Orin class GPU, yet requires further optimization. Systematic search of INT8 quantization calibration for the BiFPN fusion block weights along with structured pruning of CBAM attention maps and knowledge distillation from the YOLO11s network to a student network can result in creating a deployable network for embedded systems without significant accuracy loss.

Larger and more diverse datasets. The current experiment utilizes only one publicly available benchmark. Future research needs to evaluate the proposed architecture on larger datasets featuring images of various drones, thermal imaging, multiple drones in the scene, and drone types (fixed-wing, multirotor, micro-UAV). Cross-dataset generalization and identification of domain-specific augmentations will be key for further development and practical application.

Drones' classification and threat assessment. Once we go beyond binary detection, the next logical step in the direction of full-fledged UAV surveillance is the process of drones' classification, analyzing their movement behavior and automated assessment of threat level. For that purpose, the detection approach we proposed here offers an excellent starting point: the attention-based feature maps extracted using CBAM could be used as inputs for a classificatory head.

5.8 FINAL CONCLUSION

In this thesis, an attempt has been made to enhance the precision of UAV detection in real-time applications without any delay due to reduced speed in the inference process. The proposed approach, i.e., YOLO11s + CBAM + BiFPN, meets all the expectations by delivering 94.8% Precision, 93.7% Recall, 96.1% mAP@50, and 71.8% mAP@50–95 at 75 FPS in the case of YOLO Drone Detection Dataset [22]. The proposed architecture outperforms all the baseline models based on each evaluation metric.

The main idea underlying such excellent performance results from the fact that CBAM [14] and BiFPN [21] focus on different yet complimentary aspects of drone detection problems. In particular, attention guides the model towards drone-relevant features while feature fusion ensures the finest spatial details required for small object detection. The use of both techniques in YOLO11 framework [8] results in a very promising accuracy-speed ratio compared to all other known techniques.

These limitations — especially the lack of temporal modeling and multimodal sensing — constitute a clear set of research tasks for the future. Nevertheless, considering the limitations imposed by a single-mode visual sensor, namely an RGB camera on standard computing equipment, it has been shown that selective attention and lightweight bidirectional feature integration are extremely powerful techniques for advancing the capabilities of UAV monitoring to a new level of effectiveness.